EDGEWOOD CHEMICAL BIOLOGICAL CENTER

ECBC-TR-922

MULTIVARIABLE AND MULTIGROUP RECEIVER OPERATING CHARACTERISTICS CURVE ANALYSES FOR QUALITATIVE AND QUANTITATIVE ANALYSIS

Waleed M. Maswadeh
A. Peter Snyder

RESEARCH AND TECHNOLOGY DIRECTORATE

January 2012

U.S. ARMY
RDECOM

Disclaimer

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorizing documents.

REPORT DOCUMENTATION PAGE		Form Approved OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to everage 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering end maintaining the deta needed, end completing and reviewing this collection of information. Send comments regarding this burden estimate or any other espect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Devis Highwey, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE (DD-MM-YYYY) 06-01-2012	2. REPORT TYPE Final	3. DATES COVERED (From - To) Dec 2009 - Aug 2010

4. TITLE AND SUBTITLE Multivariable and Multigroup Receiver Operating Characteristics Curve Analyses for Qualitative and Quantitative Analysis	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) Maswadeh, Waleed M.; and Snyder, A. Peter	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Director, ECBC, ATTN: RDCB-DRD-P, APG, MD 21010-5424	8. PERFORMING ORGANIZATION REPORT NUMBER ECBC-TR-922

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT: An algorithm was developed using univariate statistics to reduce and analyze multivariate and multiple group data sets. The algorithm features the quantitative and selectivity figures of merit of receiver operating characteristics (ROC) curve methodology. This "merging" of two separate statistical analysis techniques resulted in the ability to address more than one variable in more than two experimental groups in a systematic fashion. The classic Fisher iris flower data set is treated as one variable and two cases at a time following conventional ROC curve methodology. Redundant, noisy, and low information-containing variables are removed. The remaining information-rich variables are systematically merged using ROC curve techniques. The new algorithm using ROC curve techniques produces a "master" vector of downselected variables. The ROC curve technique can be used to process any data distribution whether linear or nonlinear; the inherent trend and fundamental nature of the raw data are not compromised. No data set normalization or scaling procedures are necessary. Combining qualitative and quantitative aspects of data analysis into a univariate statistical method provides advantages in terms of algorithm understanding for the layman as well as enhanced computer efficiency and information-rich analysis.

15. SUBJECT TERMS

Multivariate analysis	Sepal	*Iris setosa*	Fisher iris flower data set
Univariate analysis	Petal	*Iris virginica*	Area under the curve
Frequency distribution	AUC	*Iris versicolor*	Angle of rotation
Receiver operating characteristics	ROC curve		

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON Renu B. Rastogi
a. REPORT U	b. ABSTRACT U	c. THIS PAGE U	UL	48	19b. TELEPHONE NUMBER (include area code) (410) 436-7545

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

20120113009

Blank

PREFACE

The work described in this report was started in December 2009 and completed in August 2010.

The use of either trade or manufacturers' names in this report does not constitute an official endorsement of any commercial products. This report may not be cited for purposes of advertisement.

This report has been approved for public release. Registered users should request additional copies from the Defense Technical Information Center; unregistered users should direct such requests to the National Technical Information Service.

Acknowledgment

Blank

CONTENTS

Blank

MULTIVARIABLE AND MULTIGROUP RECEIVER OPERATING CHARACTERISTICS CURVE ANALYSES FOR QUALITATIVE AND QUANTITATIVE ANALYSIS

1.   INTRODUCTION

The extensive amount of data generated from a set of experiments can seem daunting as one approaches the sample determination and interpretation phases. It is desirable to extract as much qualitative and quantitative information as possible from an experimental analysis. Usually, replicate analyses are mandatory for consideration of the error inputs associated with an experimental set of data. Experimental designs that, for example, rely on spectroscopy, spectrometry, or chemical shift responses can easily generate an overwhelming amount of information for a typical experiment. When experimental data from replicate analyses are combined with a suite of different sample groups for comparison purposes, visual interpretations yield relatively poor conclusions for decision-making purposes.

During the data analysis and interpretation phases, it is important to extract as much information as possible. Presentation of the raw data into a finished product requires the use of careful, well-thought-out statistical procedures. Thus, the data reduction phase is an important step in data analysis that provides a critical bridge between the raw data and the interpretation and decision-making processes.

Multivariate analysis is an attractive data reduction technique that is used to convert an extensive amount of experimental sample data into a highly reduced set of qualitative, visual information. A typical data record can contain many hundreds to thousands of variables such as wavelength, wavenumber, mass to charge ratio (m/z), retention time, or chemical shift. By converting experimental records into points, a visual two- or three-dimensional plot can be obtained that consists of a dispersion of points in which each point is a complete experimental record.[1-4] Depending on the data set and type of input, the principal component, discriminant, canonical variate, and dendrogram data reduction analyses can be implemented. The data records are reduced using the principal of a linear combination of variables. In general, most data sets do not follow or exhibit linear behavior. Nevertheless, multivariate data analysis is a widely used technique for mathematically forcing nonlinear data sets into a linear model. This can cause distortion of the data set during the data reduction analysis phase. In addition, condensing many hundreds of variables (dimensions) that are resident in a multivariate-dimensional data space into a two- or three-dimensional plot inherently produces a distorted view of the relative positioning of the experimental data points in the original multidimensional data space.

Multivariate data analysis provides qualitative accounting for a set of experiments including interpretation and decision-making. A database can be constructed from a known set of substances to characterize and possibly identify an unknown or suspect sample with respect to its presence in the database. These are practical and important objectives for qualitative data analysis.[1-4]

Another objective in data interpretation and decision-making concerns the quantitative information component. Characterization and identification of the data set with

specific examples are important tasks, but the reliability, sensitivity, and specificity of the technique are just as important. The decision-making process relies on these figures of merit.

Receiver operating characteristics (ROC) curve analysis was developed in the field of statistical decision theory and was broadened in the 1950s to the field of signal detection theory as a means of enabling radar operators to distinguish between enemy targets, friendly forces, and noise.[5,6] ROC curves report on the quantitative aspects of a data set in an analysis.[7-10] Replicate data are essential for an ROC analysis, because the replicate information contains various sources of error that inherently affect the analysis reliability and experimental error. These figures of merit factor into the decision-making process in a plot of sensitivity [true positive (TP)] on the ordinate and selectivity or specificity $(1 - TN = FP)$ on the abscissa, where TN is true negative and FP is false positive. TP and TN characterize the data set and the reliability of decisions that are made on the experimental data set. An important parameter in ROC curve analysis is the area under the curve (AUC). The AUC is a separation measurement[11-15] between two groups of interest (positive/negative, healthy/diseased, go/no go, control/test subjects, present/absent, Group A/Group B, or green/red, to name a few).

The AUC can also be considered as a measure of how well a variable can distinguish between two diagnostic groups. The AUC directly translates into the TP and false negative (FN) parameters, which are fundamental to decisions or conclusions from the data set observations and responses, including sample discrimination and/or future directions of analysis. Each point on an ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold. A cutoff or threshold value is merely a perpendicular line drawn on a standard frequency plot of two distributions.[16] A series of perpendicular lines are drawn throughout the two frequency distributions, and the lines represent sensitivity and selectivity pairs of points to characterize the control and sample distributions of data.

An ROC curve analysis retains the integrity of the data throughout the analysis. That is, unlike multivariate data analysis, a linear or nonlinear data set retains its integrity in the ROC analysis. When several ROC curves are compared, the AUC is usually the best discriminator.[11-15] The AUC is calculated by the extended trapezoidal rule.[13,14] ROC curve analysis is a univariate technique, and multivariate data analysis typically uses many variables (variates); hence, the term multivariate analysis.

Scurfield[17] offered an extensive suite of experiments for arriving at an analysis of the volume under the surface (two variables and their frequency) or hypervolume under the volume (manifold, using three variables and their frequency). Hundreds of experiments were required to delineate the boundaries and internal data space of the volumes to produce an overall accounting of the response. Li and Fine[18] took the entire data set of Scurfield[17] and used bootstrap inference probability estimation methods to statistically reduce the data set for a more manageable data analysis algorithm. Yiannoutsos et al.[19] used biochemical procedures that provided three classes of medical outcomes for human immunodeficiency virus patients. The three classes were partitioned over the same single variable, and each partition region was treated separately. A volume was calculated by plotting the (x, y, z) coordinates of the frequency probabilities from the response distribution of the one variable for the three classes.[19] The volume under the surface of

the responses (probability) was deemed equivalent to the AUC of a typical ROC curve. This analysis is unique to three classes being partitioned by one variable.

We have developed an algorithm that applies ROC curve analysis on the intensity distribution for each variable in an experimental record over an entire data set of experiments. In Part 1 of the analysis, a frequency versus intensity distribution is constructed independently for each variable. Instead of the AUC, we used the area between the curve and the diagonal (ACD) line. Variables displaying a relatively high ACD were retained and variables with relatively low ACD values were considered as noise or as not relevant for discrimination purposes. The operation of Part 1 produced a reduced set of variables that provided information relevant to the samples. In Part 2, a series of interrogations was undertaken in which two variables at a time were used to plot their (x, y) intensity pairs throughout all the experiments (cases). A vector with a given angle from the abscissa was drawn, a frequency distribution was produced, an ROC curve was constructed from the frequency distribution, and the ACD was noted. The angle of the vector was incremented, and a new ROC curve was constructed. This occurred in increments over 360°. The angle with the highest ACD value and its vector were retained for those two variables. The vector for the first two variables (V1,2) became the next independent axis (abscissa), and the third variable intensity (V3) formed the independent axis (ordinate). The cases resident on V1,2 became the x values for the y values from the respective cases on V3 (variable 3 ordinate). The ROC curve analysis was continued to produce a V1–3 vector (vectors 1, 2, and 3) and corresponding angle 1–3 (angles 1, 2, and 3). The process was repeated for every variable and resulted in (vector, angle) pairs in which the total number of pairs was equal to the number of variables minus 1. This database can therefore be used to investigate an unknown or target experimental output such as the probability of a spectrum belonging to a sample reference database.

A univariate statistical technique has been created that combines the fundamental characteristics of multivariate data analysis and the quantitative information of an ROC curve.

2.   THEORY

Details of the reduction of variables are presented in the Discussion section. The mathematical steps of the multivariable, multigroup ROC curve approach are presented as follows. Appendix A contains all tables, and Appendix B contains all figures.

2.1   Step 1

A point plot of the responses for the first two variables (V1 and V2) in Table 1 is constructed. A vector labeled V1,2 is incrementally rotated between 0 and 360° through the four-quadrant point plot. At each increment, the angle of the vector (α) is noted with respect to the origin. At angular increments, a frequency distribution of the points is made with respect to the vector acting as the abscissa. The first vector is the original abscissa, labeled 0° rotation. From the frequency distribution, an ROC curve is constructed,[7-15] and the ACD is calculated. The vector angle is incremented to a new α, and a new frequency distribution is constructed with respect to that vector's new angle placement. An ROC curve is plotted from this new frequency

distribution, and the ACD is noted. This occurs for every angular increment about the origin. Equations 1 and 2 provide the mathematical details:

$$(V1,2)_i = (V1_i^2 + V2_i^2)^{0.5}(\cos(\alpha_0 - \alpha)) \tag{1}$$

where $\quad \alpha_0 = \cos^{-1}(V1_i/(V1_i^2 + V2_i^2)^{0.5}); i = 1, 2, 3, ..., m$ (rows) $\tag{2}$

2.2 Step 2

The α and vector where the ACD is at a maximum are denoted $\alpha_{1,2}$ and V1,2 for that pair of variables:

$$(V1,2)_i = (V1_i^2 + V2_i^2)^{0.5}(\cos(\alpha_0 - \alpha_{1,2})) \tag{3}$$

2.3 Step 3

Steps 1 and 2 are repeated using the new combined column V1,2 from Step 2 with the next column V3 (see Table 1) to form vector V1–3.

2.4 Step 4

Steps 1 through 3 are repeated until all the columns are combined.

This approach reduces the original data matrix (Table 1) into one vector that combines the data from the four variables for all replicate measurements. The approach also identifies the maximum delineation and probability of separation between two groups of responses.

3. RESULTS AND DISCUSSION

3.1 Reduction of Variables

The multivariate ROC curve method is illustrated using the classic Fisher data set.[20] This data set consists of the sepal and petal widths and lengths of 50 individual flowers from *Iris setosa*, *I. versicolor*, and *I. virginica*. The three iris species constitute three separate groups, and the sepal and petal widths and lengths represent four distinct variables. The Fisher data set is commonly used as a primer or example for multivariate data analysis presentations. However, the analysis herein follows the ROC curve procedure, which uses the response of one variable to discriminate between two groups. This procedure is strengthened by the systematic adding or "merging" of all of the variable responses. Part 1 provides an analysis for arriving at the fewest number of variables containing the greatest amount of information.

10

3.1.1   *I. versicolor* and *I. virginica*

Table 1 presents the original experimental data collected by Fisher[20] and includes the lengths and widths of the sepals and petals of the three iris flower species. The analysis began by establishing the integrity of the responses of an individual variable with respect to an ROC curve determination. The sepal lengths of *I. versicolor* and *I. virginica* were plotted in two frequency distributions for an ROC curve determination on the degree of separation between the two species. Figure 1 provides two histograms (frequency distributions) of the sepal lengths of the *I. versicolor* (filled circles) and *I. virginica* (triangles) species from Table 1. There were a total of 50 points for each species; therefore, adding the filled circle or triangle ordinate values each yields 50.

A typical ROC curve analysis consists of moving a vertical threshold or cutoff line from the left to the right (as in Figure 1) in increments (the increment here was 3.3 mm on the abscissa). In Figure 1, six vertical lines, or thresholds, are presented. A plot of the ROC curve is simply a representation of what percentage or fraction of each group resides on and to the left of a selected line. Usually, the left-hand distribution is chosen as the starting point of the analysis. The fraction of points on and to the left of line 1 consisted of 8% (4/50 = 0.08) *I. versicolor* (filled circles) and 2% (1/50 = 0.02) *I. virginica* (triangles). The fraction of each group was determined with respect to its own distribution. Thus, there was no meaning in the addition of the two percentages (pair of points, top of Figure 1). Lines 2–6 followed suit as *I. versicolor*, *I. virginica*: 42%, 6%; 78%, 38%; 98%, 76%; 100%, 88%; and 100%, 100%. The six points were marked in Figure 2 and provided the basis for the ROC curve. The *I. versicolor* and *I. virginica* species were plotted as fractions on the ordinate and abscissa, respectively. The analysis was with respect to the left-hand distribution; therefore, the left-hand distribution took the label of TP on the ordinate. The right-hand distribution took the form of the abscissa, or $1 - TN = FP$. The ROC curve was obtained as in Figure 2, and the points corresponding to each threshold in Figure 1 were marked accordingly. The ACD, which is the area between the curve and the 45° diagonal line in Figure 2, could be calculated. Considering that the entire square space enclosed by 0, 0; 0, 1; 1, 0; and 1, 1 had an area equal to 1.0, the diagonal line provided two regions of 0.5 each.

3.1.2   The ACD

The ACD is characteristic of the degree of separation between the two species (groups). When a tested variable provides essentially no discrimination ability between two groups, a Cartesian plot of $(1 - TN, TP)$ yields an experimental line close to the 45° line. This occurs when the dispersion of variable responses in the frequency plot for the two groups overlaps to a significant extent; the ACD approaches zero. When a variable provides a significant degree of separation, a plot of $(1 - TN, TP)$ yields an exponential curve that starts at 0, 0, continues nearly straight up to approach 0, 1, and then is almost horizontal to 1, 1. The ACD in this case approaches 0.5, and this signifies a high degree of separation, or specificity, between the two measured groups. This occurs when the frequency distribution of variable responses for two groups has a high degree of separation with very little to no overlap.

In Figure 2, the ACD is 0.2896. By multiplying the ACD by 200, a percent degree of separation was obtained. Thus, $0.2896 \times 200 = 57.9$ or ~58%. Therefore, the sepal length has a 58% probability of distinguishing between the two species of iris, which is a relatively poor degree of separation. This ROC curve process was repeated for the sepal width between the two iris species. An ACD of 0.1636 was obtained, and the percent separation was equal to 32.7%. Figures 3 and 4 present the frequency distribution plot and the ROC curve, respectively. The ROC curve was closer to the diagonal line in Figure 4 as compared to Figure 2. The sepal width displayed approximately half the degree of separation (32.7%) with respect to the sepal length (58%) for the Fisher data set.

The analysis was repeated for the petal lengths and widths for the two species. Table 2 presents the ACD and percent separation for the analysis of the four variables. Figure 5 is a plot of the four ACD values in Table 2. Columns 1–4 in Figure 5 represent the sepal length, sepal width, petal length, and petal width, respectively, for each of the three species in Table 1. Table 2 and Figure 5 provide significant information, because the results show that in Fisher's original data, the sepal contribution did not provide as good discrimination ability as the petal information. This result was attained by use of simple univariate ROC curve statistics as compared to the classical multivariate data analysis treatment.[1–4]

Both the petal length and width allowed for a 96% degree of separation between the dispersion of the two measurements. Therefore, the original four variables may be reduced to only one variable, i.e., either petal length or petal width, for a satisfactory degree of separation between the two iris species.

3.1.3   *I. setosa* and *I. virginica*

The above process was repeated in a comparison of the four iris variables between the *I. setosa* and *I. virginica* species. Table 3 provides the reduced data set, and Figure 6 is a plot of the ACD values from Table 3. Note that petal lengths and widths, separately, provided a 100% degree of separation, and both produced an ideal ROC curve with an ACD of 0.5. This result signifies that multivariate analysis is not necessary when a simple univariate ROC curve analysis provides a satisfactory level of discrimination. Figure 7, A–D includes frequency distribution plots of the four variables for the *I. setosa* (open circles) and *I. virginica* (triangles) species with information for the ROC curves shown in the figure insets. The petal information provided a 100% degree of separation for the two species, whereas the sepal length data produced a very high degree of separation (96.9%). Sepal width information resulted in a relatively poor degree of separation (66.8%).

3.1.4   *I. setosa* and *I. versicolor*

Univariate iteration was performed for a comparison of the *I. setosa* (open circles) and *I. versicolor* (triangles) species as shown in Figure 8, A–E and Table 4. Data reduction and analysis were performed in a similar fashion as shown in Figures 6 and 7, A–D and Tables 2 and 3. Again, the petal width and length variables achieved a 100% separation between the two species. For the entire data set, there was no need for further analysis, because the three iris species could be distinguished from one another with a 96 to 100% degree of separation where only one variable was necessary for the distribution of any two groups.

3.1.5   Iris Flower Analysis

The above data analysis essentially identifies variables that contribute a significant amount of discriminating information. Generally, experimental data consist of many variables (hundreds to thousands) such as spectroscopy (wavenumber or wavelength) and spectrometry (m/z or ion mobility drift time) spectra. It is these types of data sets that would most benefit by a reduction of variables. The variable reduction also identifies the least discriminating and noisy variables that need not be considered in subsequent analyses. Also, further analysis is necessary when a reduction of variables procedure provides multiple variables that yield less than satisfactory discrimination capability.

3.2   Data Integration

3.2.1   Variable Processing

The iris data set did not require further analysis. However, to show the strengths of the data analysis concept presented herein, all four variables were considered for further processing, and the resulting three different sets of species analyses were analyzed and compared.

3.2.2   Data Space Point Rotation

Part 2 of the data analysis consisted of merging the variables in a systematic fashion with an ROC curve analysis at each variable inclusion. This accounting of the variables took place with two groups of experiments at a time. In the case of the iris data, analysis of three species required three sets of two groups: *I. setosa, I. versicolor*; *I. setosa, I. virginica*; and *I. versicolor, I. virginica*. Each group was treated separately with all four variables in a systematic procedure that used univariate statistics. The *I. versicolor, I. virginica* pair of species was addressed first. The figure symbols are as presented above.

3.2.2.1   Sepal Variables, V1,2

Instead of a frequency plot of distance for only one variable, an (x, y) pair of axes was constructed in which the abscissa was that of variable 1 (V1, sepal length), and the ordinate was that of variable 2 (V2, sepal width). Figure 9A shows the point plot for the *I. versicolor* and *I. virginica* species. Fifty points were plotted for each species. A series of steps was performed in either of two equivalent ways, and one of the procedures is presented in a comprehensive fashion.

The data space of points was uniformly rotated in 10° increments, and an analysis was performed at each angle increment including an ROC curve ACD determination. At the start, no rotation was necessary, and this was labeled as a 0° rotation (Figure 9A). A perpendicular line was drawn to the abscissa from all 100 points. This was the same as the original x axis or V1 axis, where V1 is a label for the first variable vector or sepal length. Intensity bins were formed on the x axis, and the number of filled circle and triangle points in each bin were summed and plotted in Figure 9B as a frequency distribution. The two sets of

13

points yielded two distribution curves. An ROC curve analysis was performed on the two distributions as shown in Figure 9C to calculate an ACD. The abscissa, which consisted of the placements of the sepal length of all 100 points, the 0° rotation value, and the ACD, was then saved.

The 100 points were then uniformly rotated 10° (Figure 9D). A perpendicular line from all 100 points was drawn to the x axis; this can be considered a modified x axis from the original abscissa. Intensity bins were formed on this new x axis, and the numbers of filled circle and triangle points were counted and plotted (Figure 9E) as a frequency distribution. The two sets of points formed two distribution curves as in Figure 9E, and the ROC curve is shown in Figure 9F. The ACD is noted with the 10° modified x axis, with the placement of all 100 points on that axis.

These steps were repeated at every 10° increment of data space point rotation. An ROC curve ACD was derived for each pair of distributions of the sepal distances at each rotation increment. Figure 9G shows a 330° rotation of the points, which produced the maximum ACD, and Figure 9H shows the distribution of both sets of points for the *I. versicolor* and *I. virginica* species from Figure 9G. Figure 9I is the ROC curve for the data in Figure 9G.

Figure 10A is a plot of a select set of the ROC curves at different data space rotations. Note that there are ROC curves that lie below the 45° line. Between 0° and 180°, the 1 − TN and TP values were plotted with respect to the *I. versicolor* points lying to the left of the *I. virginica* species in a V1–V2 plot. Between 180° and 360°, the points were rotated such that the *I. versicolor* points were to the right of the *I. virginica* species. This produced ROC curves below the diagonal line. Figure 10B is a plot of the rotation angle versus its respective ROC curve ACD value. The 330° rotation provided the maximum ACD value of 0.29. An ACD of 0.29 is equivalent to a 58% degree of separation for the two iris species using both sepal dimensions. An ACD of 0.29 is not a satisfactory degree of separation; rather, an ACD close to 0.5 is desired. Therefore, it was necessary to consider the next variable for inclusion and merging to possibly improve the ACD experimental value.

3.2.2.2   Sepal Variables and Petal Length, V1–3

The vector generated by the angle at 330° was labeled V1,2, because it provided the best separation between the two distributions with respect to the ACD value with the sepal dimensions. V1,2 became the new abscissa, and the ordinate represented the vector for the third variable, i.e., petal length or V3. Figure 11A presents a data space with the abscissa and ordinate as V1,2 and V3, respectively. All 100 points were plotted in the data space accordingly. Note that each of the 100 points had its x value on the V1,2 axis. All steps and procedures were performed with a 10° rotation increment along with an ROC curve analysis for the *I. versicolor* and *I. virginica* species data. Figure 11B shows the point plot at a 250° rotation of the data points, which yielded a maximum ACD value, and Figure 11C shows an overlay of the two frequency distribution curves. Figure 11D presents an ROC curve analysis, and Figure 11E is a plot that shows the angle of rotation versus the ACD. Note that the maximum ACD of 0.49 occurred at a 250° angle. An ACD of 0.49 is equivalent to a 98% degree of separation for the two iris species. Therefore, the petal length (V3) provides a major source of differentiation compared to both sepal variables.

14

3.2.2.3 All Four Iris Variables, V1–4

The vector at 250° was referred to as V1–3, and it contained information from vectors 1, 2, and 3. This vector became the new abscissa, and variable 4 or petal width became V4 on the ordinate. All 100 points were plotted in the data space (Figure 12A). This representation translated a four-dimensional data set (four variables or vectors) into a two-dimensional data space without loss of the inherent data set characteristics and trends. Figure 12B shows a 330° rotation of the dispersion of points, and Figure 12C presents an overlay of the two frequency distribution curves. A rotation angle of 330° provided the highest ACD upon an ROC curve analysis (Figure 12D) of all rotation angles. An ACD of 0.4998 was obtained, which corresponds to a 99.6% degree of separation between the two iris species.

3.2.2.4 Petal Variables, V3,4

This analysis leads to the question regarding whether all four variables are necessary to produce a 99.6% degree of separation. To answer this question, the data set was analyzed using only petal length (V3) and petal width (V4). This reduced the problem to two variables and two groups, which is still beyond the standard ROC curve analysis of only one variable distinguishing between two groups. Generally, the more information that can be applied to a statistical problem, the greater the degree that a satisfactory resolution can be achieved. Note that this new statistical technique allows for many variables to be introduced in an analysis of two groups for qualitative as well as ROC curve quantitative and selectivity information. As more variables are considered, a larger ACD and hence better discrimination between the two groups is achieved. The frequency distributions at maximum ACD values versus the numbers of variables considered can be compared in Figures 9H (V1–2), 11C (V1–3), and 12C (V1–4). The respective ACD (percent separation) values were 0.29 (58%), 0.49 (98%), and 0.498 (99.96%).

Figure 13A is a plot of petal length (V3) versus petal width (V4) for the *I. versicolor* (filled circles) and *I. virginica* (triangles) species. Figure 13B presents the two frequency distributions of the points in Figure 13A, and Figure 13C presents the ROC curve. Figure 13D shows the angle versus the ACD plot for the iris petal information; note that at a rotation of 280°, the maximum ACD value was achieved. Figure 14A shows the plot where the points in Figure 13A were rotated 280°, and Figure 14B presents its frequency distribution. Figure 14C shows the ROC curve for the 280° rotation of points. Using only the V3 and V4 variables produced an ACD of 0.495 (99% separation) compared to using all four variables, which yielded an ACD of 0.4998 or 99.6% separation (Figure 12, C and D). The difference in separation efficiency of the two species was negligible.

The entire process can be repeated for discrimination purposes for any two groups or cases consisting of any number of variables. This process can be repeated for the *I. setosa* and *I. virginica* species. However, the analysis in Part 1 established that the petal lengths and widths were sufficient to provide a 100% separation (Table 3). This was also true for the separation of the *I. setosa* and *I. versicolor* species (Table 4). These qualitative and quantitative discrimination analyses were accomplished without the use of matrix algebra as required by multivariate data analysis methods.

3.3   Alternate Method of Analysis

The above analysis has an equivalent procedure for attaining the same results. Instead of rotating the entire data set of points at every angle increment, a vector is drawn at 10° from the x axis. The points remain in their positions, and a perpendicular line is drawn from the 100 points onto the 10° vector. The vector undergoes a frequency distribution analysis and results in two frequency distributions. An ROC curve analysis takes place, and the vector, angle, and ACD are stored. The vector is then placed 20° from the x axis, and perpendicular lines are drawn to that new vector to denote the placement of the 100 points on that vector. This procedure provides the same results as the rotation of data space points.

4.   CONCLUSIONS

A univariate statistical method was presented to collect, reduce, and analyze a multivariable response for replicates of more than two cases or groups. Conventional ROC curve analysis is the backbone for the method. The new statistical univariate data analysis method herein provides an ROC curve analysis with the ability to incorporate more than one variable and more than two groups in the analysis and conclusions for qualitative differentiation and selectivity purposes. The raw data remain in its inherent trend and nature. No data set normalization or scaling procedures are necessary. The mean and standard deviation, which are fundamental values for multivariate data analysis, are not needed. By nature, the ROC curve procedures can be applied to any kind of data distribution in addition to the classic Gaussian trend, such as step functions and skewed distributions, whether they are linear or nonlinear. Multivariate data analysis, on the other hand, necessarily forces a data set into a linear model, because the algorithm relies on a linear combination of variables. The ROC curve method presented here has the ability to translate a multivariable (or multidimension or multivector) data set into a one-variable or one-dimensional response analysis while preserving the inherent nature of the distribution, whether it is linear or nonlinear.

ROC curve analyses yield results that determine which variables are best used for the critical decision-making process that distinguishes two experimental groups. The variables themselves dictate which of the entire data set will form subsets, and this provides groupings of experimental cases (spectral points in data space). The measurement vehicle for this determination is the ACD. For a multivariate dendrogram analysis, it is the distance between each case that determines which cases form the subgroups and the relative separation of each subgroup.

The new algorithm using ROC curve techniques produces a "master" vector. The master vector (reference or library vector) is a systematic integration of the chosen set of variables. The master vector can be used in a practical situation to identify an unknown sample (case), and it will be presented elsewhere. This method also preserves the TP, FN, and FP probability quantitative information of a data set with multiple species of interest.

Combining qualitative and quantitative aspects of data analysis into a univariate statistical method provides advantages in terms of algorithm understanding for the layman, computer efficiency, and information-rich analysis.

LITERATURE CITED

1.   Ryzin, V. Cluster Analysis as a Basis for Purposive Sampling of Projects in Case Study Evaluations. *Am. J. Eval.* **1995**, *16*, 109–119.

2.   McLachlan, G.J. *Discriminant Analysis and Statistical Pattern Recognition*; Wilcy-Interscience: New York, 2004.

3.   De Maesschalck, R.; Jouan-Rimbaud, D.; Massart, D.L. The Mahalanobis Distance. *Chemom. Intell. Lab. Syst.* **2000**, *50*, 1–18.

4.   Cho, J.; Gemperline, P.J. Pattern-Recognition Analysis of Near-Infrared Spectra by Robust Distance Method. *J. Chemom.* **2005**, *9*, 169–178.

5.   Detection Theory and Applications. *Proc. IEEE* **1970**, *58*, 607–852.

6.   Lusted, L.B. *Introduction to Medical Decision Making*; C. C. Thomas: Springfield, IL, 1968.

7.   Ask Professor Mean Webpage: ROC, http://www.cmh.edu/stats/ask/roc.asp (accessed December 2011).

8.   LaBaer, J. So, You Want to Look for Biomarkers (Introduction to Special Biomarkers Issue). *J. Proteome Res.* **2005**, *4*, 1053–1059.

9.   Vitzthum, F.; Behrens, F.; Anderson, N.L.; Shaw, J.H. Protcomics: From Basic Research to Diagnostic Application: A Review of Requirements and Needs. *J. Proteome Res.* **2005**, *4*, 1086–1097.

10.   Park, S.H.; Goo, J.M.; Jo, C.-H. Receiver Operating Characteristic (ROC) Curve: Practical Review for Radiologists. *Korean J. Radiol.* **2004**, *5*, 11–18.

11.   DeLong, E.R.; DeLong, D.M.; Clarke-Pearson, D.L. Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics* **1981**, *44*, 837–845.

12.   Metz, C.E. Basic Principles of ROC Analysis. *Semin. Nucl. Med.* **1978**, *8*, 283–298.

13.   Hanley, J.A.; McNeil B.J. The Meaning and Use of Area Under a Receiver Operating Characteristic (ROC) Curve. *Radiology*, **1982**, *143*, 29–36.

14.   *Numerical Recipes: The Art of Scientific Computing*, 3rd ed.; Press, W.H.; Teukolsky, S. A.; Vetterling, W.T.; Flannery, B.P., Eds.; Cambridge University Press: New York, 2007.

15.   Marzban, C. The ROC Curve and the Area Under It as Performance Measure. *Weather Forecast.* **2004**, *19*, 1106–1114.

16.   Altman, D.G. *Practical Statistics for Medical Research*, 1st ed.; Chapman and Hall: London, **1991**, *14*, 396–439.

17.   Scurfield, B.K. Multiple-Event Forced-Choice Tasks in the Theory of Signal Delectability. *J. Math. Psychol.* **1996**, *40*, 253–269.

18.   Li, J.; Fine, J.P. ROC Analysis with Multiple Classes and Multiple Tests: Methodology and Its Application in Microarray Studies. *Biostatistics* **2008**, *9*, 566–576.

19.   Yiannoutsos, C.T.; Nakasb, C.T.; Naviac, B.A. Assessing Multiple-Group Diagnostic Problems with Multi-Dimensional Receiver Operating Characteristic Surfaces: Application to Proton MR Spectroscopy (MRS) in HIV-Related Neurological Injury. *NeuroImage* **2008**, *40*, 248–255.

20.   Fisher, R.A. The Use of Multiple Measurements in Taxonomic Problems. *Ann. Eugen.* **1936**, *7*, 179–188.

GLOSSARY

ACD	area between the curve and the diagonal
AUC	area under the curve
FN	false negative
FP	false positive
m/z	mass to charge ratio
ROC	receiver operating characteristics
TN	true negative
TP	true positive
V1	vector (variable) 1
V1,2	statistical combination of vector (variable) 1 and vector (variable) 2
V1–3	statistical combination of vectors 1, 2, and 3

Blank

TABLES

Table 1. Fisher data set. The set consisted of three species of iris flowers. Replicate measurements of four variables included sepal and petal lengths and widths. Data are in millimeters.

I. setosa				*I. versicolor*				*I. virginica*			
Sepal Length	Sepal Width	Petal Length	Petal Width	Sepal Length	Sepal Width	Petal Length	Petal Width	Sepal Length	Sepal Width	Petal Length	Petal Width
50	33	14	2	65	28	46	15	64	28	56	22
46	34	14	3	62	22	45	15	67	31	56	24
46	36	10	2	59	32	48	18	63	28	51	15
51	33	17	5	61	30	46	14	69	31	51	23
55	35	13	2	60	27	51	16	65	30	52	20
48	31	16	2	56	25	39	11	65	30	55	18
52	34	14	2	57	28	45	13	58	27	51	19
49	36	14	1	63	33	47	16	68	32	59	23
44	32	13	2	70	32	47	14	62	34	54	23
50	35	16	6	64	32	45	15	77	38	67	22
44	30	13	2	61	28	40	13	67	33	57	25
47	32	16	2	55	24	38	11	76	30	66	21
48	30	14	3	54	30	45	15	49	25	45	17
51	38	16	2	58	26	40	12	67	30	52	23
48	34	19	2	55	26	44	12	59	30	51	18
50	30	16	2	50	23	33	10	63	25	50	19
50	32	12	2	67	31	44	14	64	32	53	23
43	30	11	1	56	30	45	15	79	38	64	20
58	40	12	2	58	27	41	10	67	33	57	21
51	38	19	4	60	29	45	15	77	28	67	20
49	30	14	2	57	26	35	10	63	27	49	18
51	35	14	2	57	29	42	13	72	32	60	18
50	34	16	4	49	24	33	10	61	30	49	18
46	32	14	2	56	27	42	13	61	26	56	14
57	44	15	4	57	30	42	12	64	28	56	21
50	36	14	2	66	29	46	13	62	28	48	18
54	34	15	4	52	27	39	14	77	30	61	23
52	41	15	1	60	34	45	16	63	34	56	24
55	42	14	2	50	20	35	10	58	27	51	19
49	31	15	2	55	24	37	10	72	30	58	16
54	39	17	4	58	27	39	12	71	30	59	21
50	34	15	2	62	29	43	13	64	31	55	18
44	29	14	2	59	30	42	15	60	30	48	18
47	32	13	2	60	22	40	10	63	29	56	18
46	31	15	2	67	31	47	15	77	26	69	23

Table 1, continued.

I. setosa				I. versicolor				I. virginica			
Sepal Length	Sepal Width	Petal Length	Petal Width	Sepal Length	Sepal Width	Petal Length	Petal Width	Sepal Length	Sepal Width	Petal Length	Petal Width
51	34	15	2	63	23	44	13	60	22	50	15
50	35	13	3	56	30	41	13	69	32	57	23
49	31	15	1	63	25	49	15	74	28	61	19
54	37	15	2	61	28	47	12	56	28	49	20
54	39	13	4	64	29	43	13	73	29	63	18
51	35	14	3	51	25	30	11	67	25	58	18
48	34	16	2	57	28	41	13	65	30	58	22
48	30	14	1	61	29	47	14	69	31	54	21
45	23	13	3	56	29	36	13	72	36	61	25
57	38	17	3	69	31	49	15	65	32	51	20
51	38	15	3	55	25	40	13	64	27	53	19
54	34	17	2	55	23	40	13	68	30	55	21
51	37	15	4	66	30	44	14	57	25	50	20
52	35	15	2	68	28	48	14	58	28	51	24
53	37	15	2	67	30	50	17	63	33	60	25

Table 2. ROC ACD and percent separation between the *I. versicolor* and *I. virginica* species for sepal and petal widths and lengths.

Variable	ACD	Separation, %
1 Sepal length	0.2896	57.9
2 Sepal width	0.1636	32.7
3 Petal length	0.4822	96.4
4 Petal width	0.4804	96.1

Table 3. ROC ACD and percent separation between the *I. setosa* and *I. virginica* species for the sepal and petal widths and lengths.

Variable	ACD	Separation, %
1 Sepal length	0.4846	96.9
2 Sepal width	−0.3342	66.8
3 Petal length	0.5000	100
4 Petal width	0.5000	100

Table 4. ROC ACD and percent separation between the *I. setosa* and *I. versicolor* species for the sepal and petal widths and lengths.

Variable	ACD	Separation, %
1 Sepal length	0.432600	86.5
2 Sepal width	−0.424600	84.9
3 Petal length	0.500000	100
4 Petal width	0.500000	100
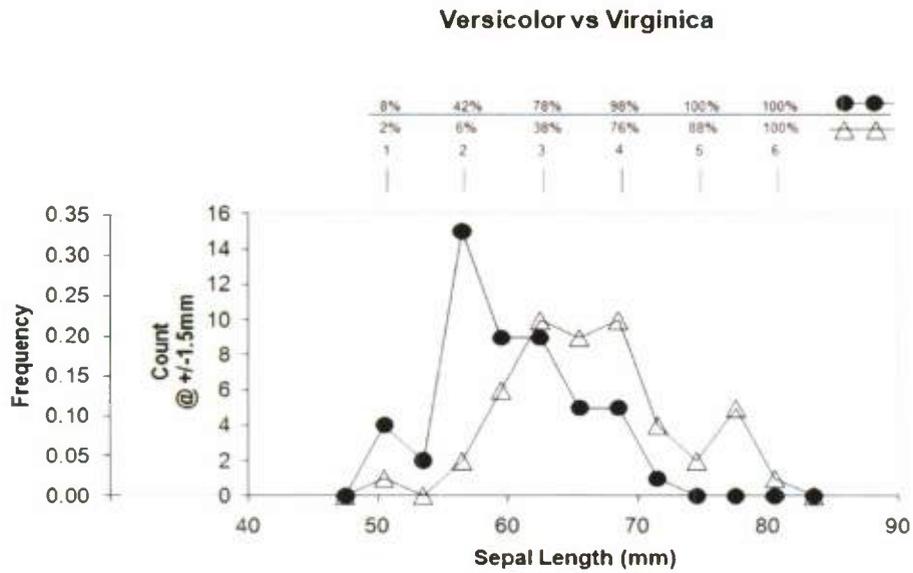
Figure 1. Frequency plots of the sepal lengths of 50 separate *I. versicolor* (filled circles) and 50 *I. virginica* (triangles) flowers. Data are from the Fisher data set.[20]
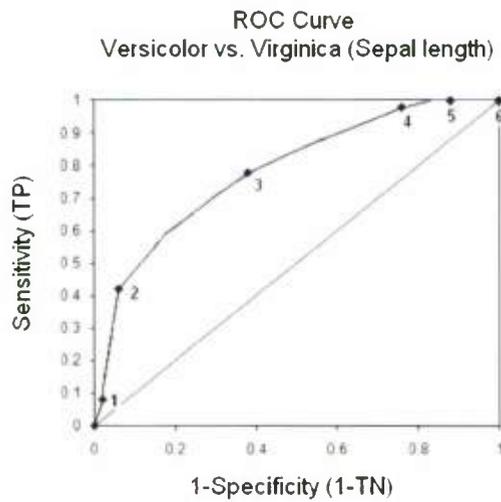
Figure 2. ROC curve of the Figure 1 data set.

Versicolor vs Virginica
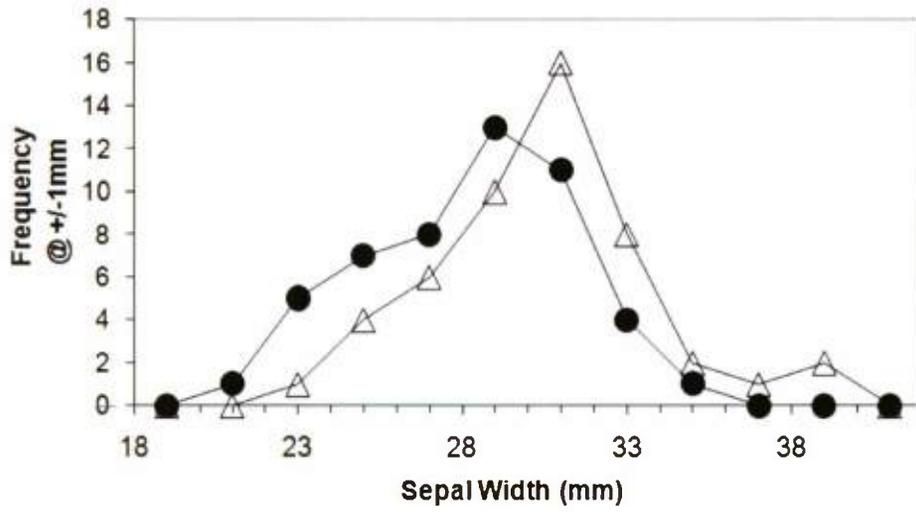
Figure 3. Frequency plots of the sepal widths of 50 separate *I. versicolor* (filled circles) and 50 *I. virginica* (triangles) flowers. Data are from the Fisher data set.[20]
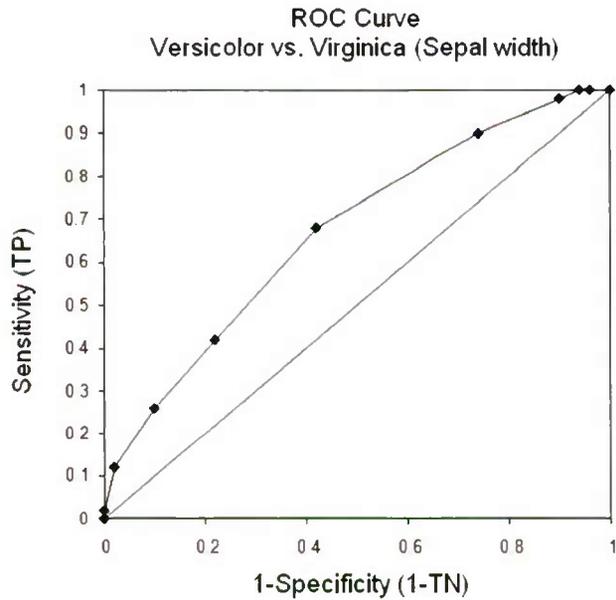
Figure 4. ROC curve of the data shown in Figure 3.
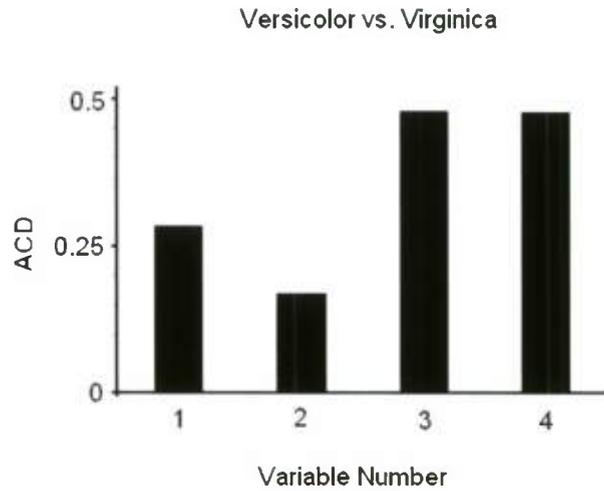
Figure 5. Histogram of the data in Table 2. Columns 1–4 represent sepal length, sepal width, petal length, and petal width, respectively.
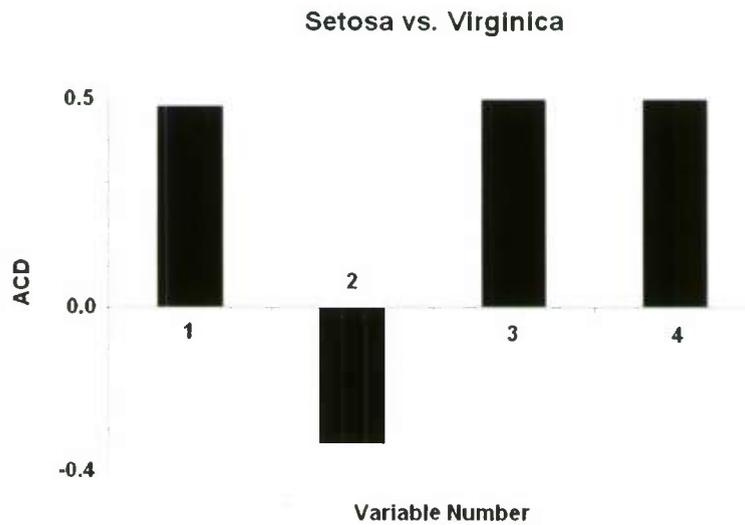
Figure 6. Histogram of the data in Table 3. Columns 1–4 represent sepal length, sepal width, petal length, and petal width, respectively.

Setosa vs Virginica
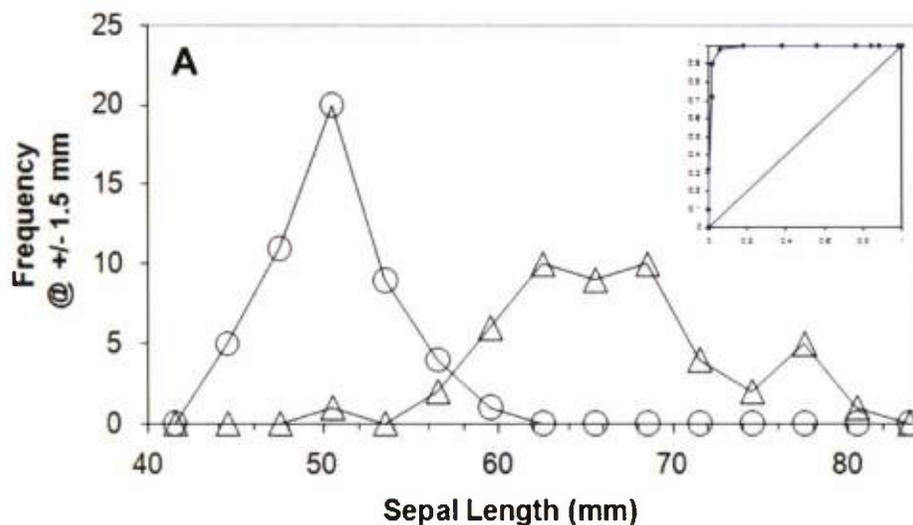
Figure 7A. *I. setosa* (open circles) and *I. virginica* (triangles) frequency distribution plots for sepal length for 50 replicates of each species. Inset displays ROC curve for the analysis.
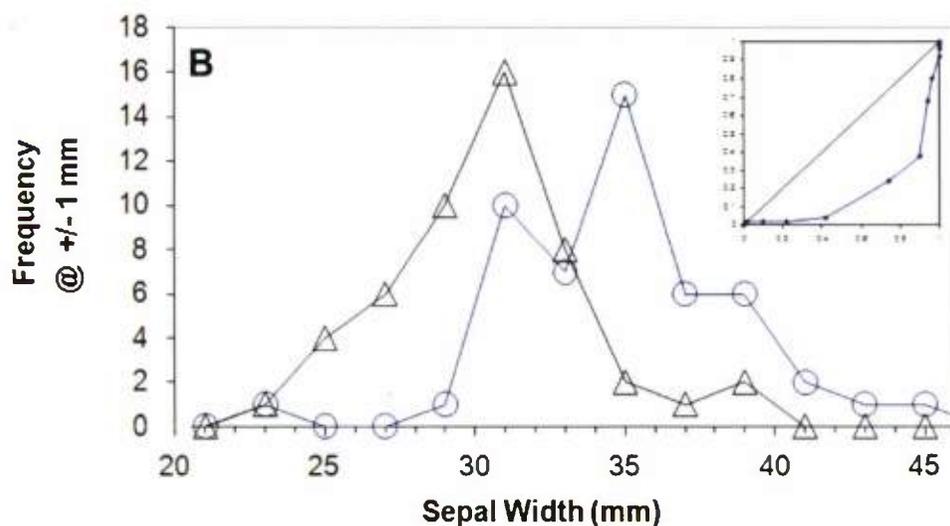
Figure 7B. *I. setosa* (open circles) and *I. virginica* (triangles) frequency distribution plots for sepal width for 50 replicates of each species. Inset displays ROC curve for the analysis.
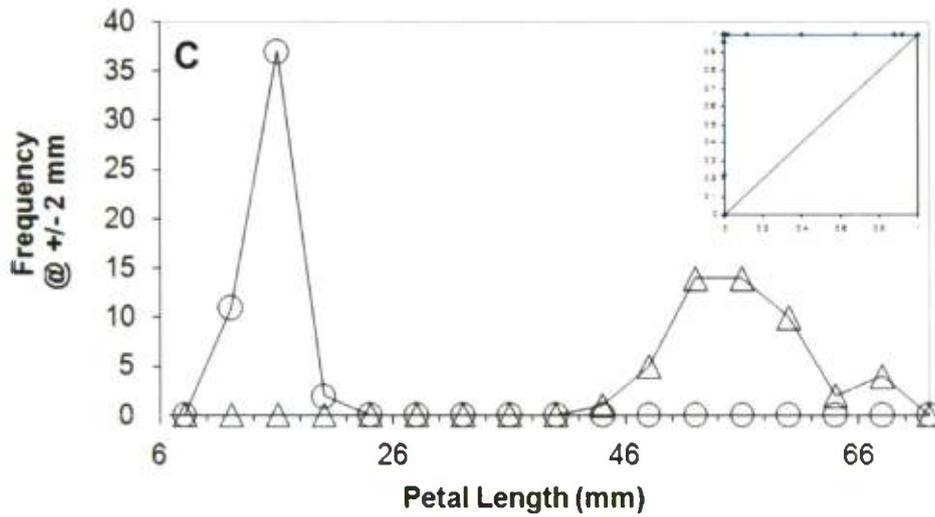
Figure 7C. *I. setosa* (open circles) and *I. virginica* (triangles) frequency distribution plots for petal length for 50 replicates of each species. Inset displays ROC curve for the analysis.
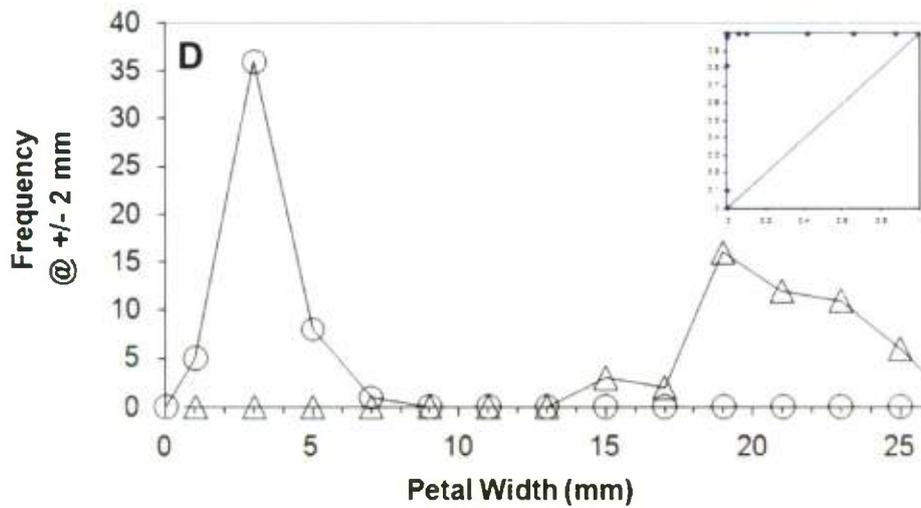
Figure 7D. *I. setosa* (open circles) and *I. virginica* (triangles) frequency distribution plots for petal width for 50 replicates of each species. Inset displays ROC curve for the analysis.

Setosa vs. Versicolor
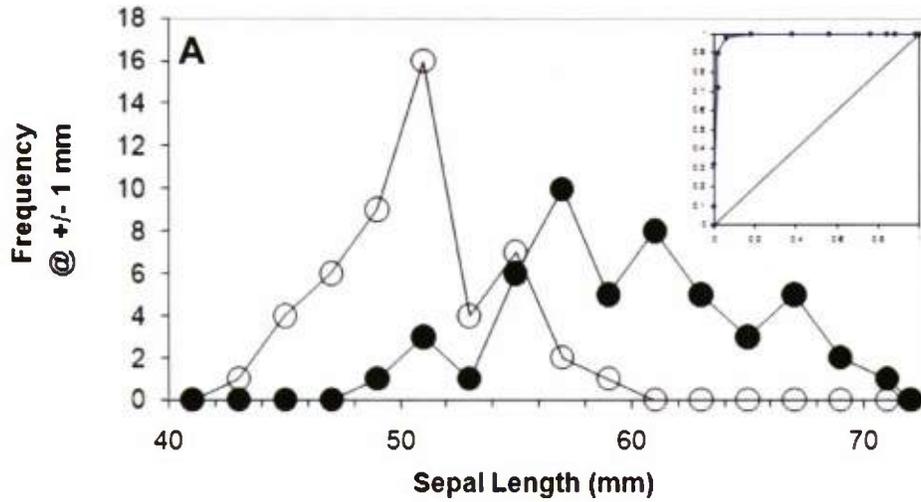
Figure 8A. *I. setosa* (open circles) and *I. versicolor* (closed circles) frequency distribution plot for sepal length for 50 replicates of each species. Inset displays ROC curve for the analysis.
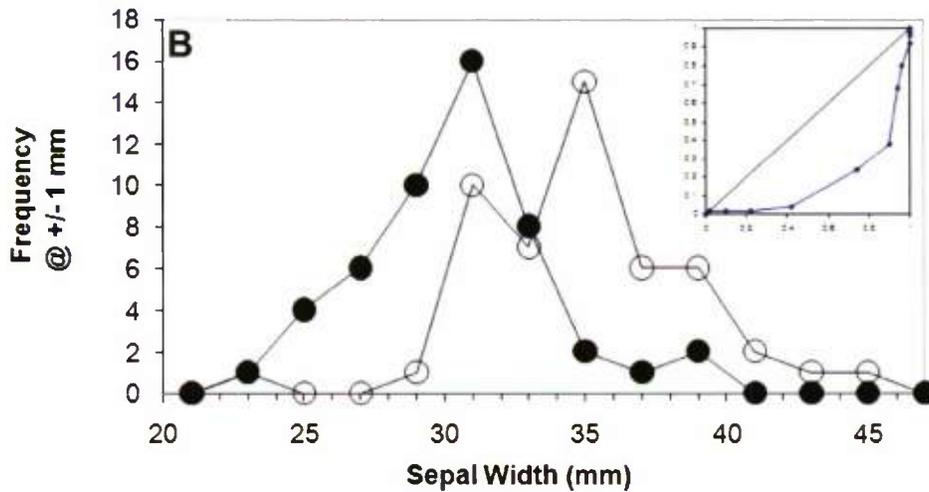
Figure 8B. *I. setosa* (open circles) and *I. versicolor* (closed circles) frequency distribution plot for sepal width for 50 replicates of each species. Inset displays ROC curve for the analysis.
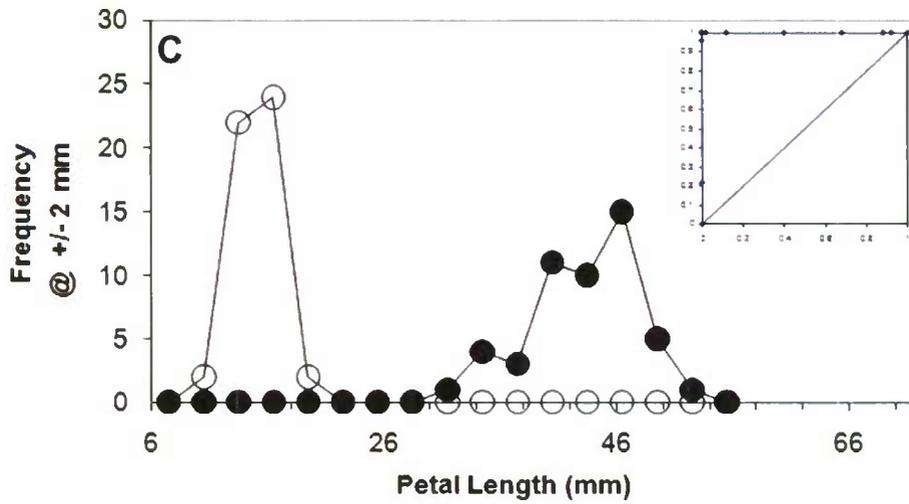
Figure 8C. *I. setosa* (open circles) and *I. versicolor* (closed circles) frequency distribution plot for petal length for 50 replicates of each species. Inset displays ROC curve for the analysis.
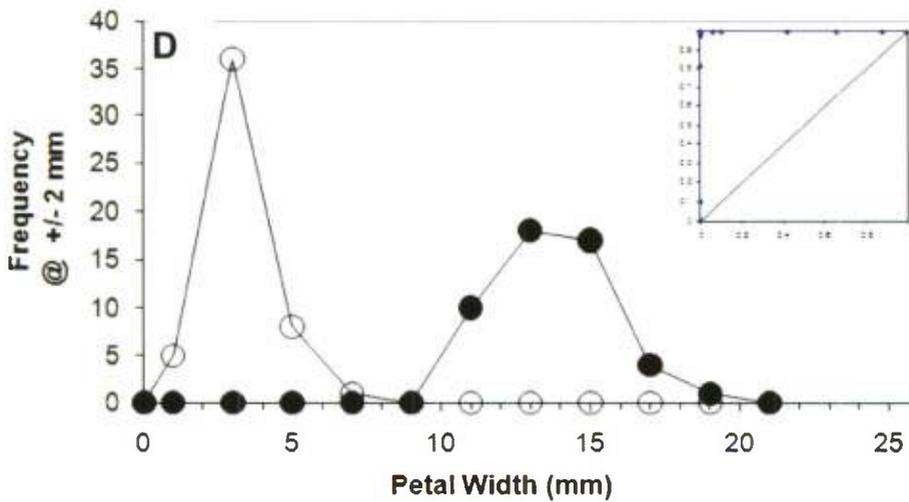
Figure 8D. *I. setosa* (open circles) and *I. versicolor* (closed circles) frequency distribution plots for petal width for 50 replicates of each species. Inset displays ROC curve for the analysis.

Setosa vs. Versicolor
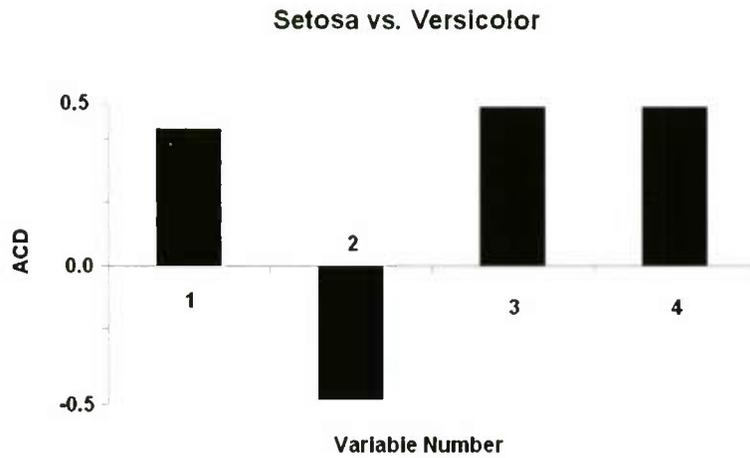
Figure 8E. Histogram of the iris measurement variables versus ACD data in Table 4. Columns 1–4 represent sepal length, sepal width, petal length, and petal width, respectively.

Versicolor vs. Virginica @ 0 degree angle
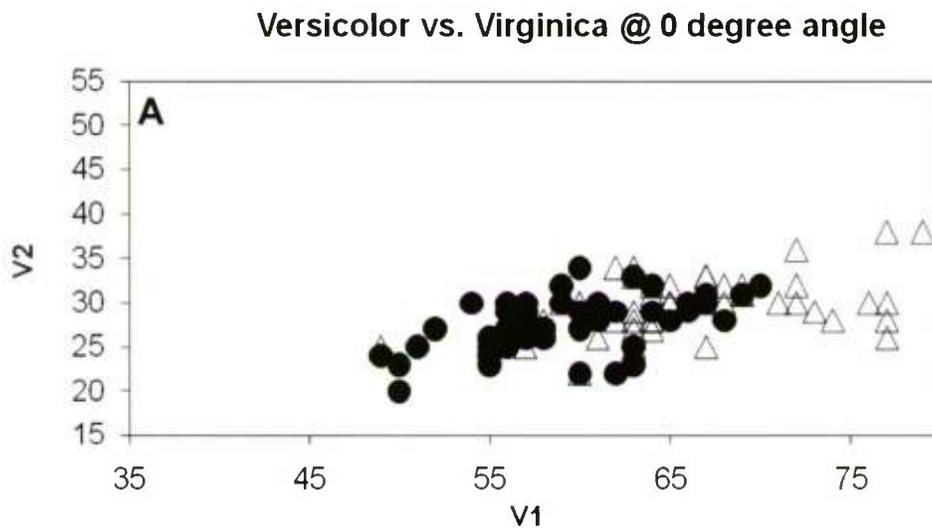

Figure 9A. Sepal length (V1) versus sepal width (V2) point plot distribution of the 50 replicates for both *I. versicolor* (filled circles) and *I. virginica* (triangles) at a 10° rotation of the points.
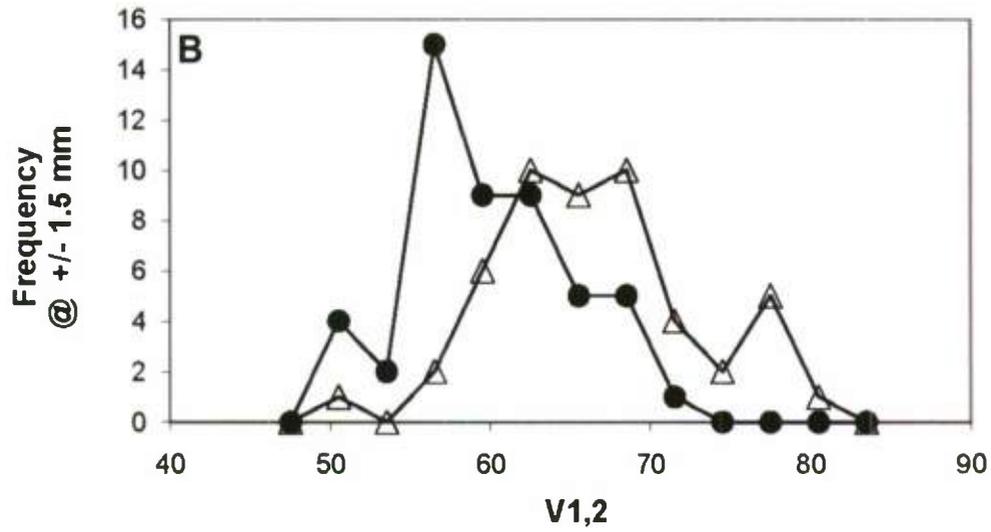
Figure 9B. Frequency distribution of points shown in Figure 9A.

Versicolor vs. Virginica @ 0 degree angle
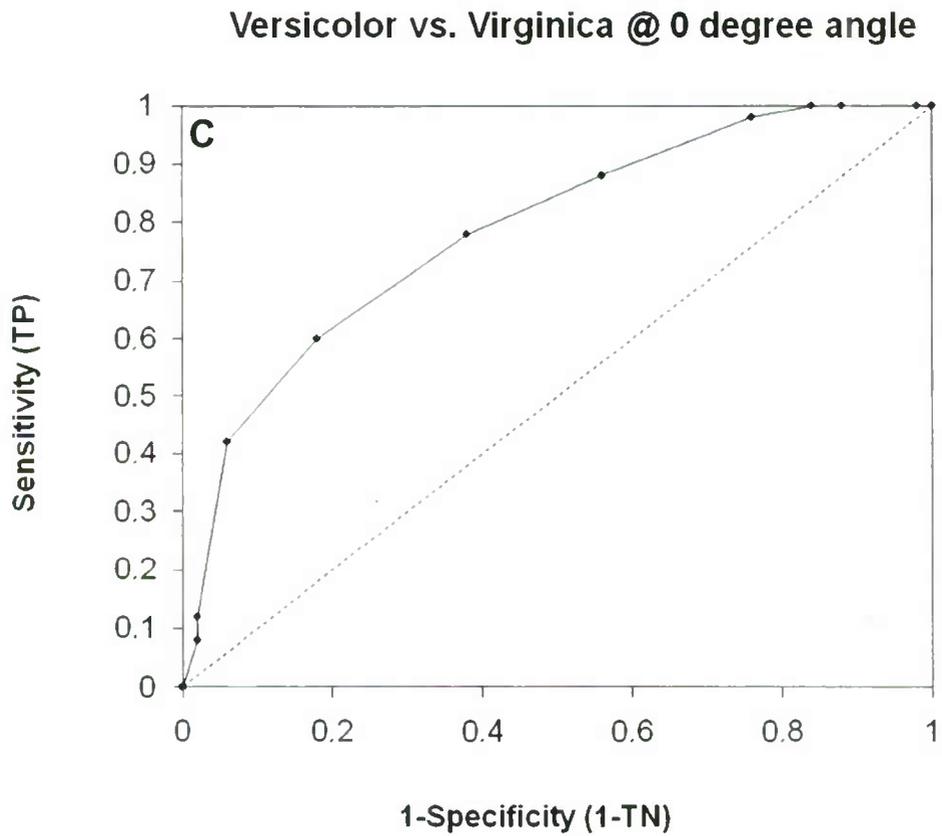
Figure 9C. ROC curve of the frequency distribution shown in Figure 9B.

Versicolor vs. Virginica @ 10 degree angle
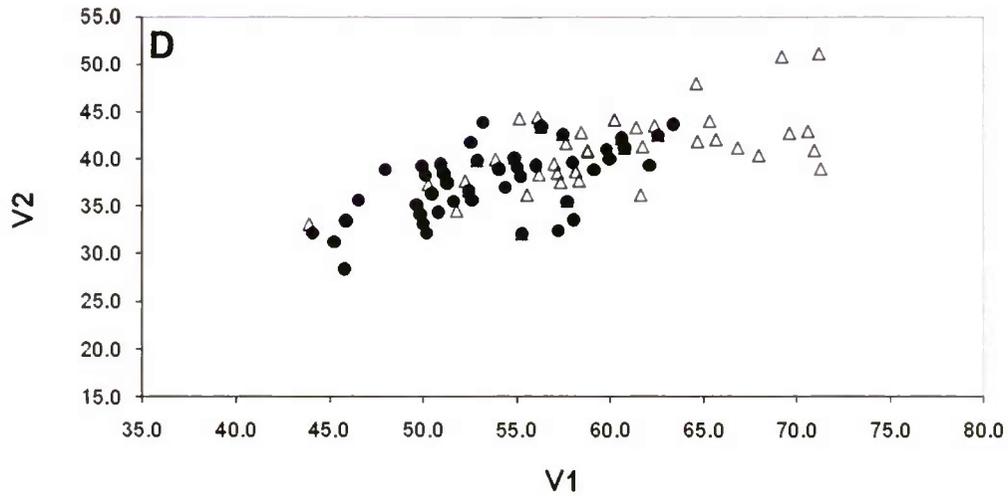
Figure 9D. Sepal length (V1) versus sepal width (V2) point plot distribution of the 50 replicates for both *I. versicolor* (filled circles) and *I. virginica* (triangles) at a 10° rotation of the points.
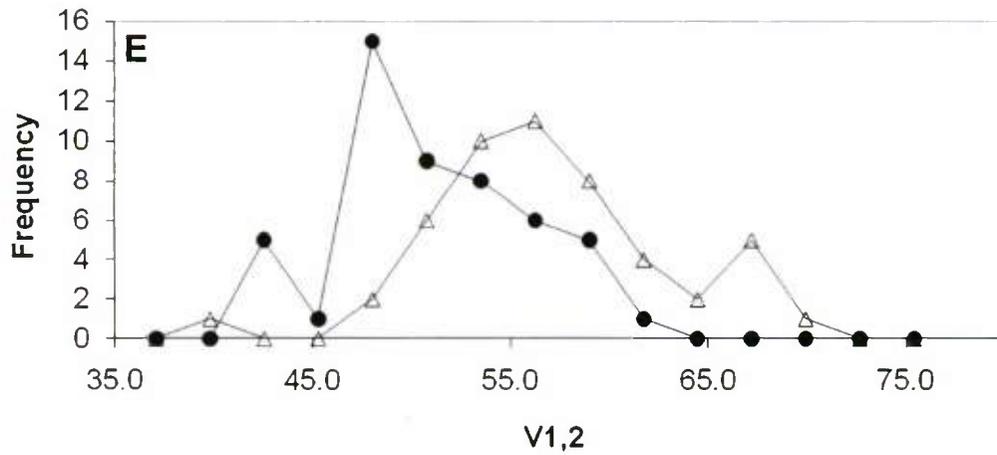
Figure 9E. Frequency distribution of points shown in Figure 9D.

Versicolor vs. Virginica @ 10 degree angle
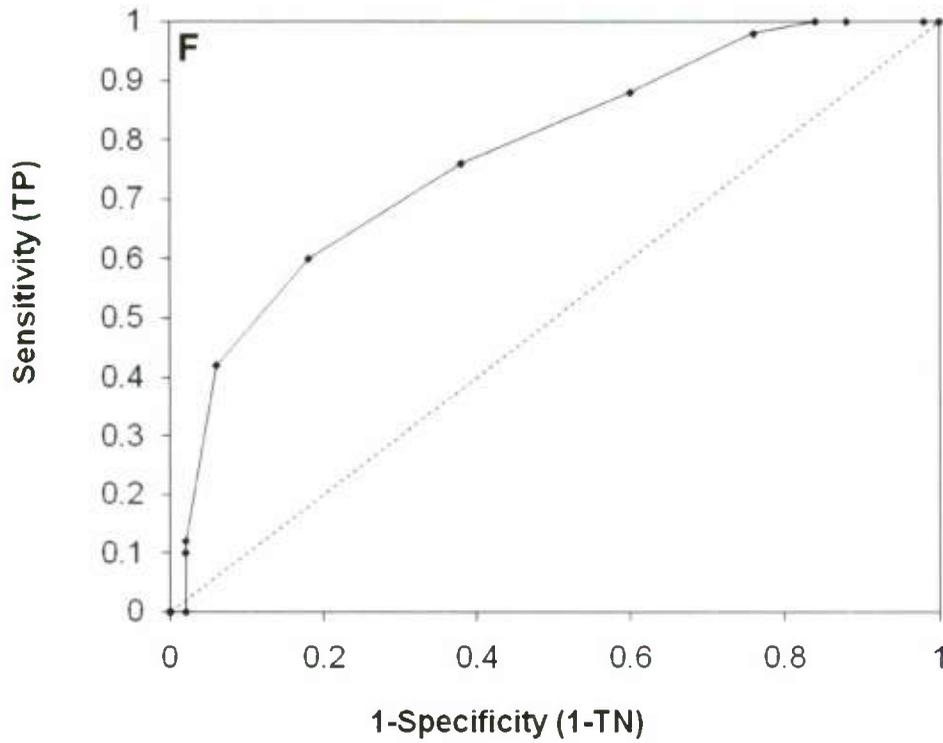
Figure 9F. ROC curve of the frequency distribution shown in Figure 9E.

Versicolor vs. Virginica @ 330 degree angle
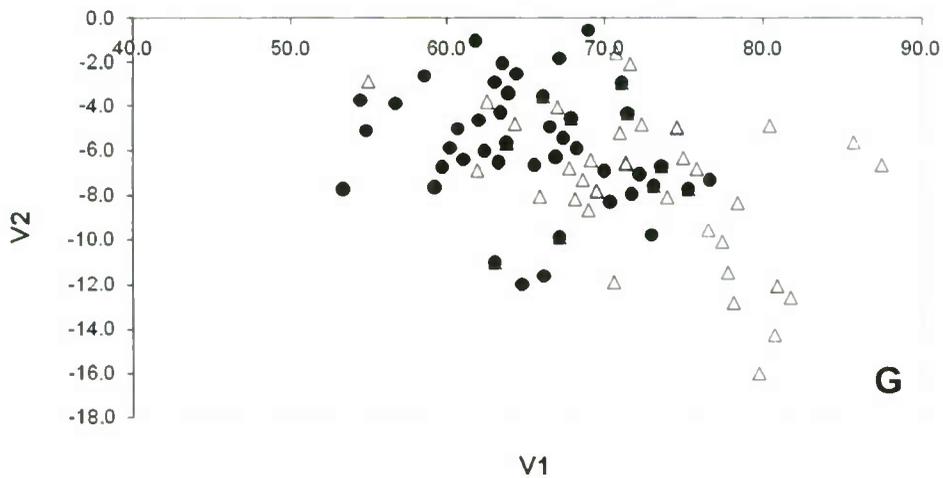

Figure 9G. Sepal length (V1) versus sepal width (V2) point plot distribution of the 50 replicates for both *I. versicolor* (filled circles) and *I. virginica* (triangles) at a 10° rotation of the points.

Versicolor vs. Virginica @ 330 degree angle
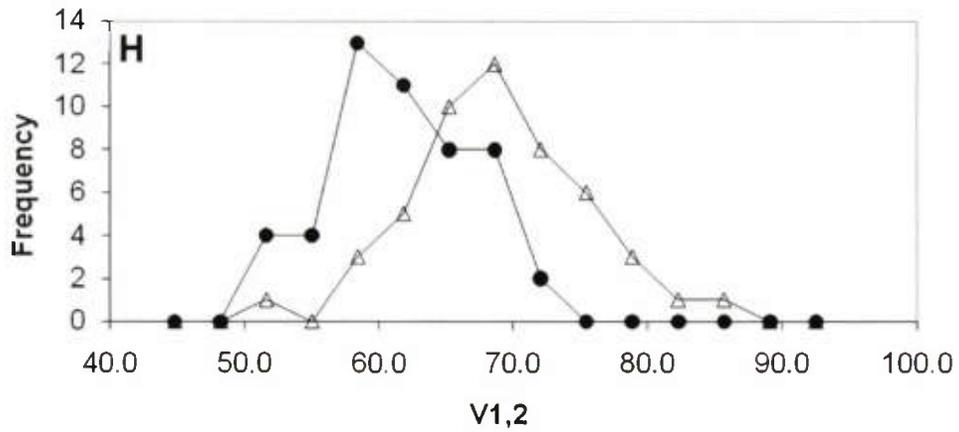
Figure 9H. Frequency distribution of points shown in Figure 9G.

Versicolor vs. Virginica @ 330 degree angle
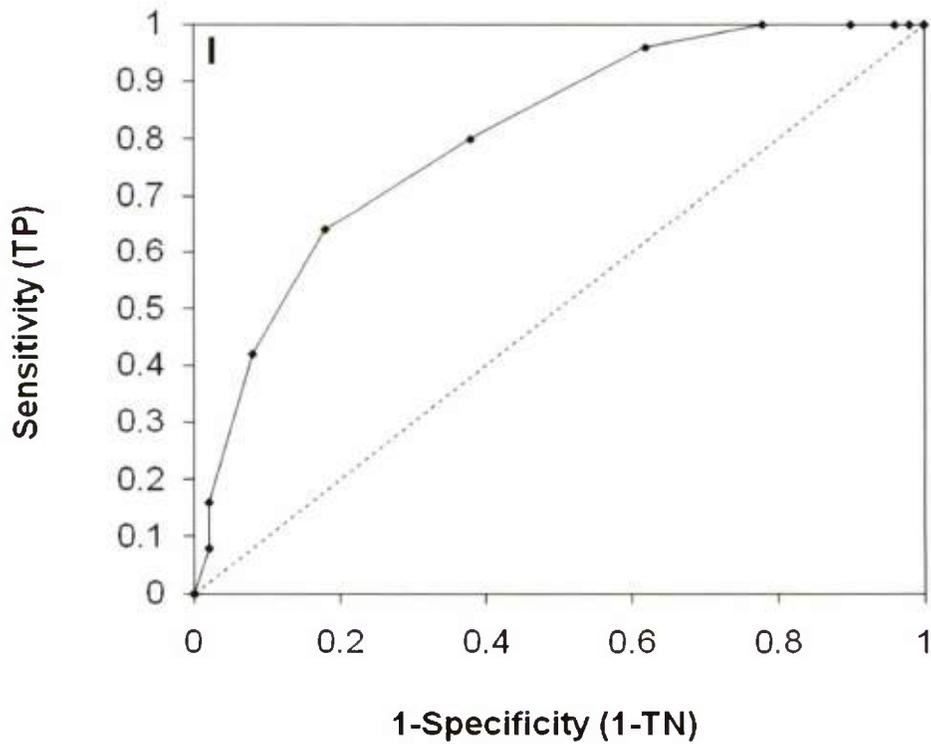
Figure 9I. ROC curve of the frequency distribution in Figure 9G.

Versicolor vs. Virginica from 0-360 degrees @ 20 degree increments
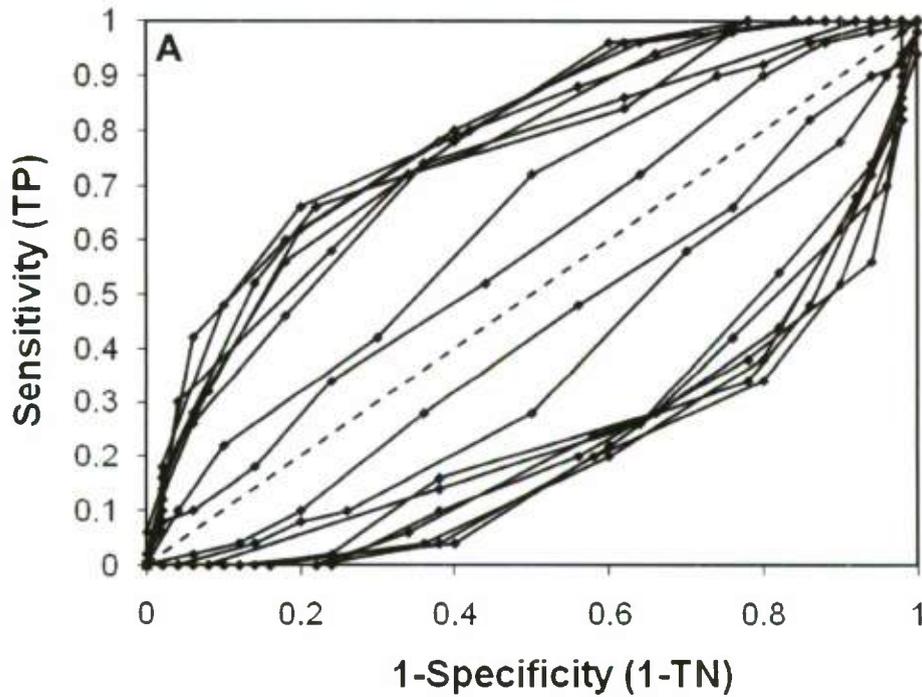
Figure 10A. Plot of the ROC curves between 0° and 360° rotation of the point distribution in Figure 9D in 20° increments.

ACD vs. Rotated Angle (V1,2)
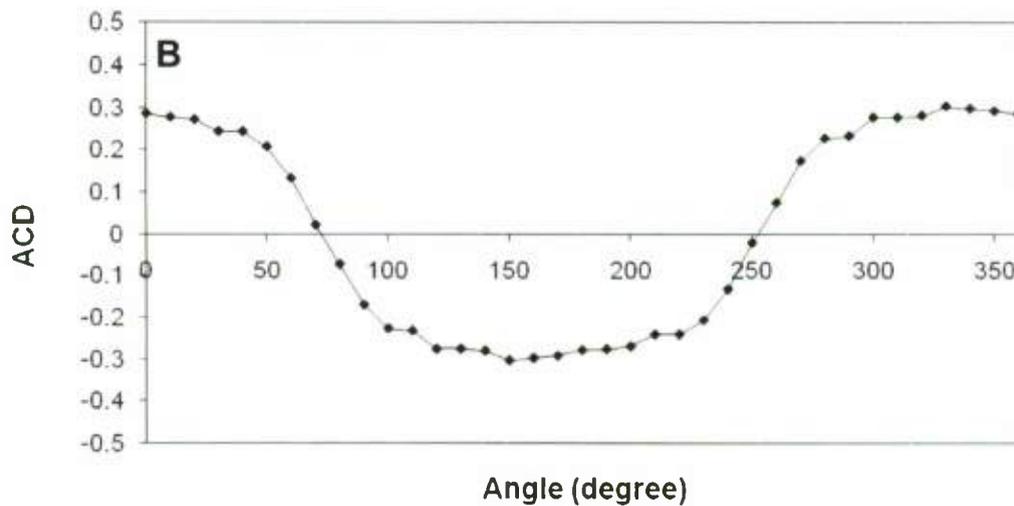
Figure 10B. Plot of angle of rotation of points in Figure 9D versus ACD.

Versicolor vs. Virginica @ 0 degree angle
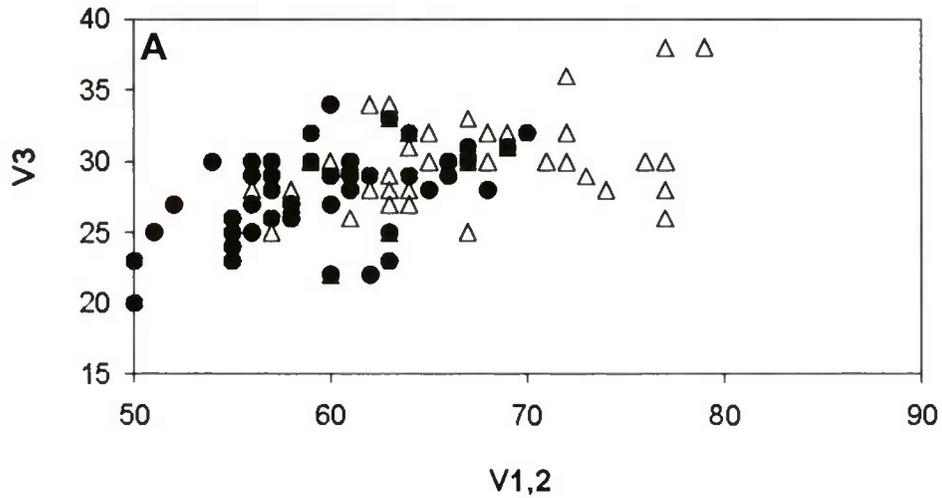
Figure 11A. Sepal length, sepal width (V1,2) versus petal length (V3) point plot distribution of the 50 replicates for both *I. versicolor* (filled circles) and *I. virginica* (triangles) with no rotation of the points.


Versicolor vs. Virginica @ 250 degree angle
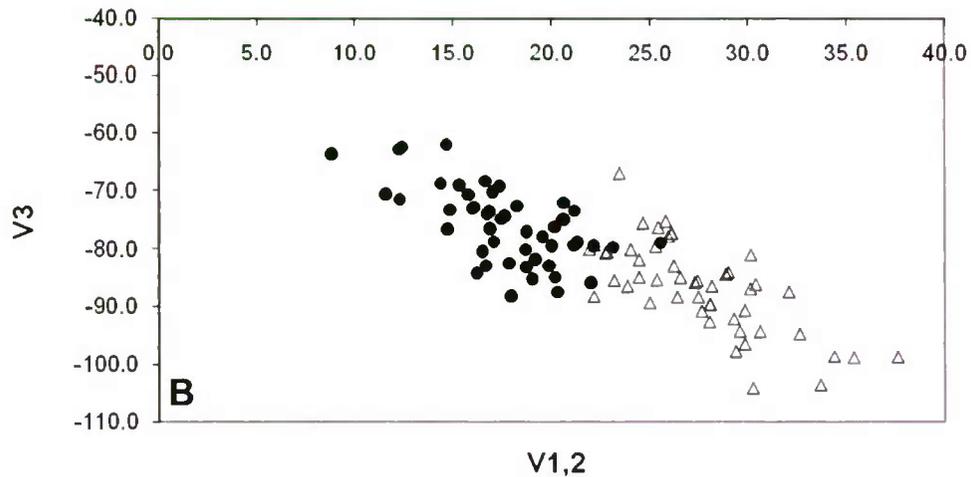
Figure 11B. Sepal length, sepal width (V1,2) versus petal length (V3) point plot distribution of the 50 replicates for both *I. versicolor* (filled circles) and *I. virginica* (triangles) at a 250° rotation.

Versicolor vs. Virginica @ 250 degree angle
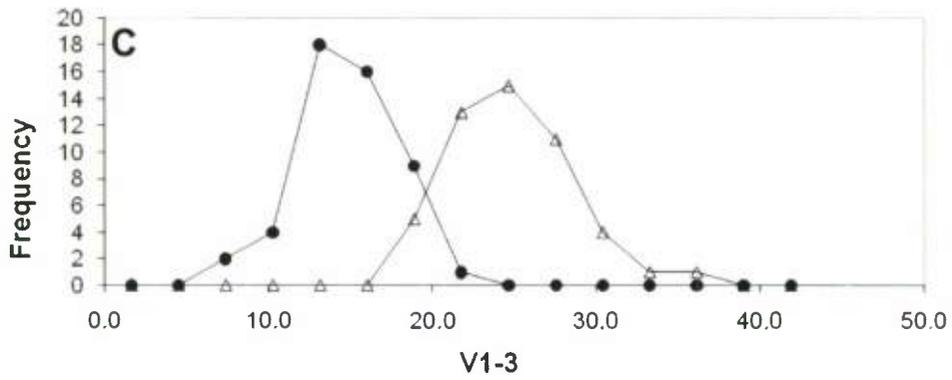
Figure 11C. Frequency distribution of the data in Figure 11B.

Versicolor vs. Virginica , V1-3 @ 250 degree angle
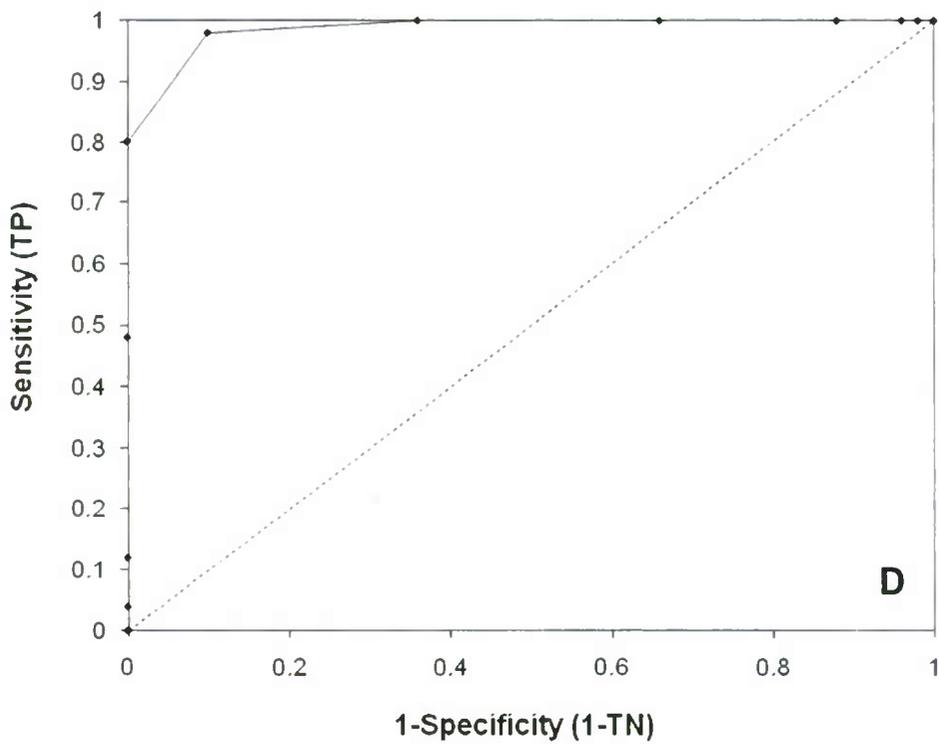
Figure 11D. ROC curve plot of the frequency distribution in Figure 11C.

ACD vs. Rotated Angle (V1-3)
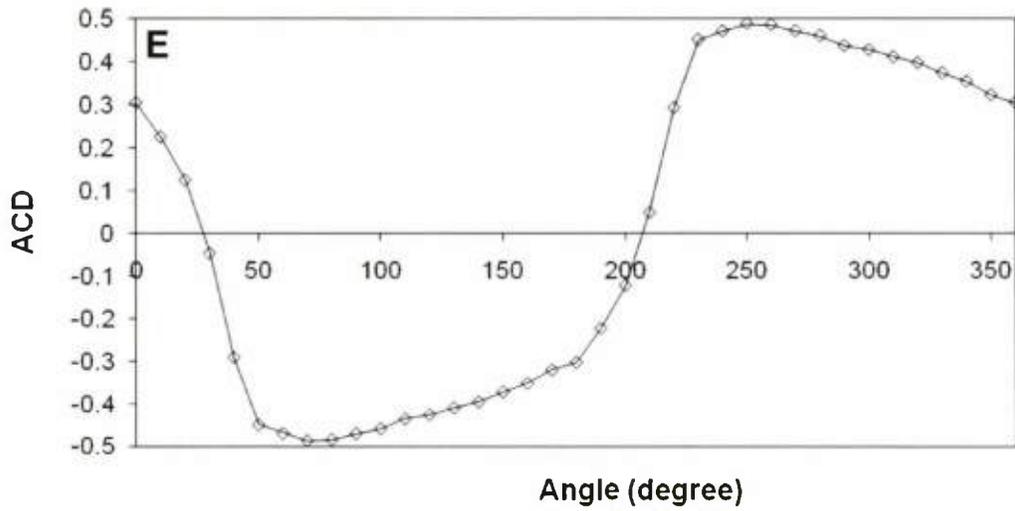
Figure 11E. Plot of angle of rotation of points versus ACD (E).

## Versicolor vs. Virginica @ 0 degree angle
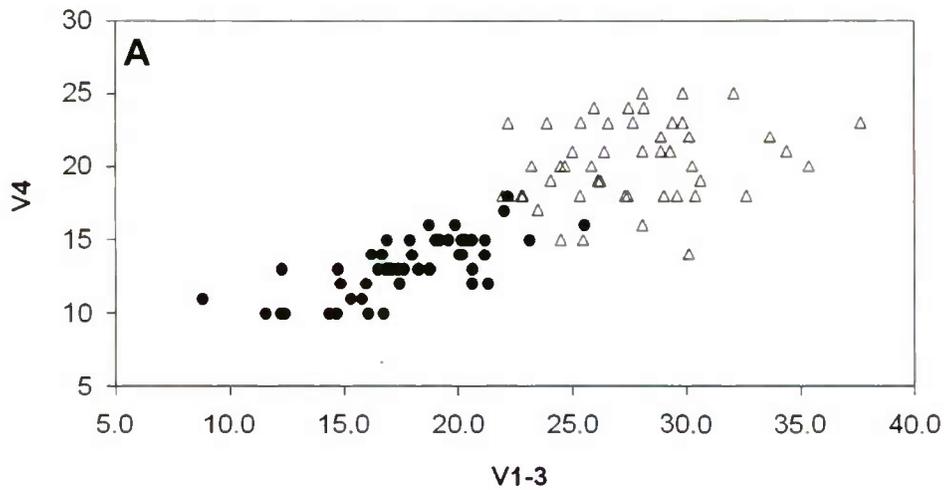
Figure 12A. Sepal length, sepal width, petal length (V1–3) versus petal width (V4) point plot distribution of the 50 replicates for both *I. versicolor* (filled circles) and *I. virginica* (triangles) with no rotation of the points.

Versicolor vs. Virginica @ 330 degree angle
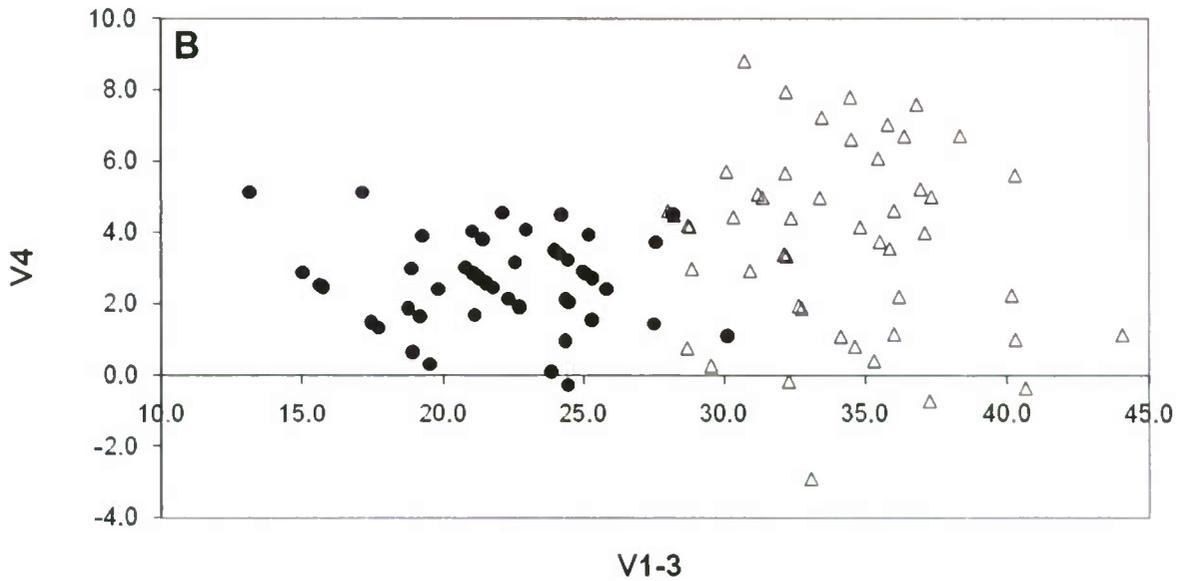
Figure 12B. Sepal length, sepal width, petal length (V1–3) versus petal width (V4) point plot distribution of the 50 replicates for both *I. versicolor* (filled circles) and *I. virginica* (triangles) at a 330° rotation of the points.
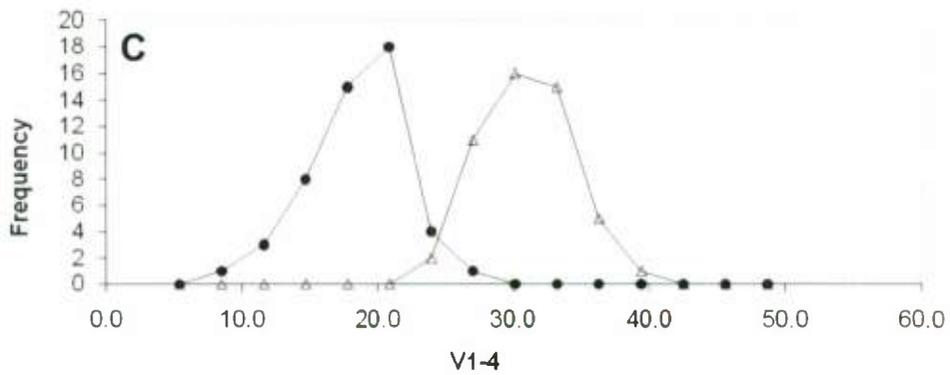
Figure 12C. Frequency distribution at a 330° rotation of points.

Versicolor vs. Virginica @ 330 degree angle
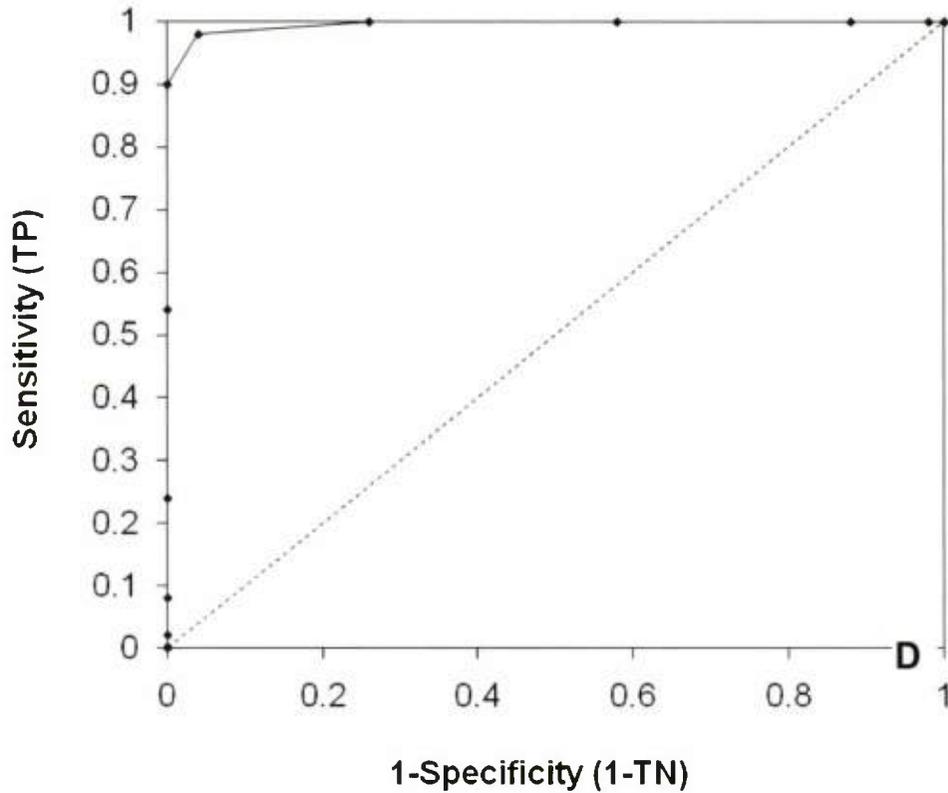
Figure 12D. ROC curve plot of the data shown in Figure 12C.

Versicolor vs. Virginica @ 0 degree angle
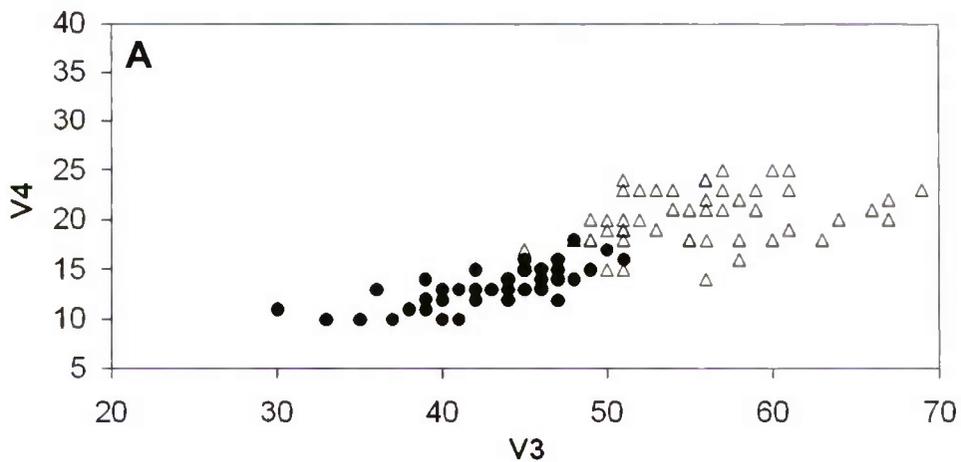
Figure 13A. Petal length (V3) versus petal width (V4) point plot distribution of the 50 replicates for both *I. versicolor* (filled circles) and *I. virginica* (triangles) with no rotation of the points.
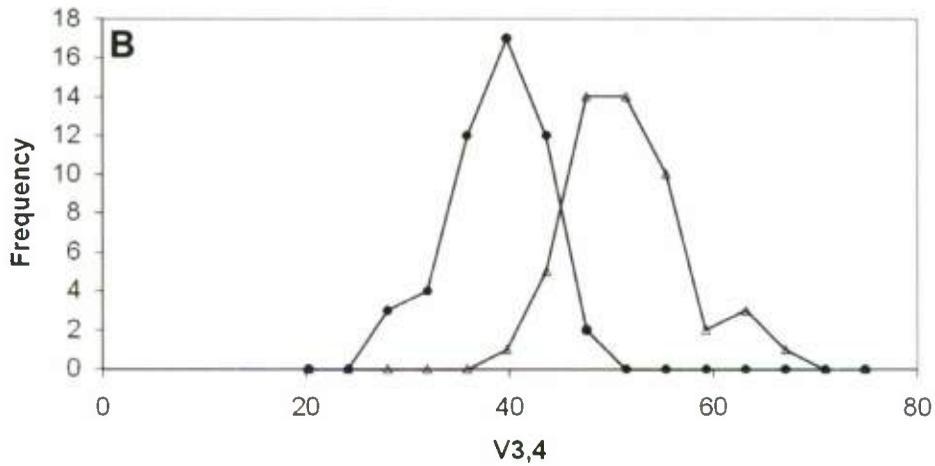
Figure 13B. Petal length (V3) versus petal width (V4) for both *I. versicolor* (filled circles) and *I. virginica* (triangles): frequency distribution at 0° rotation of the points.
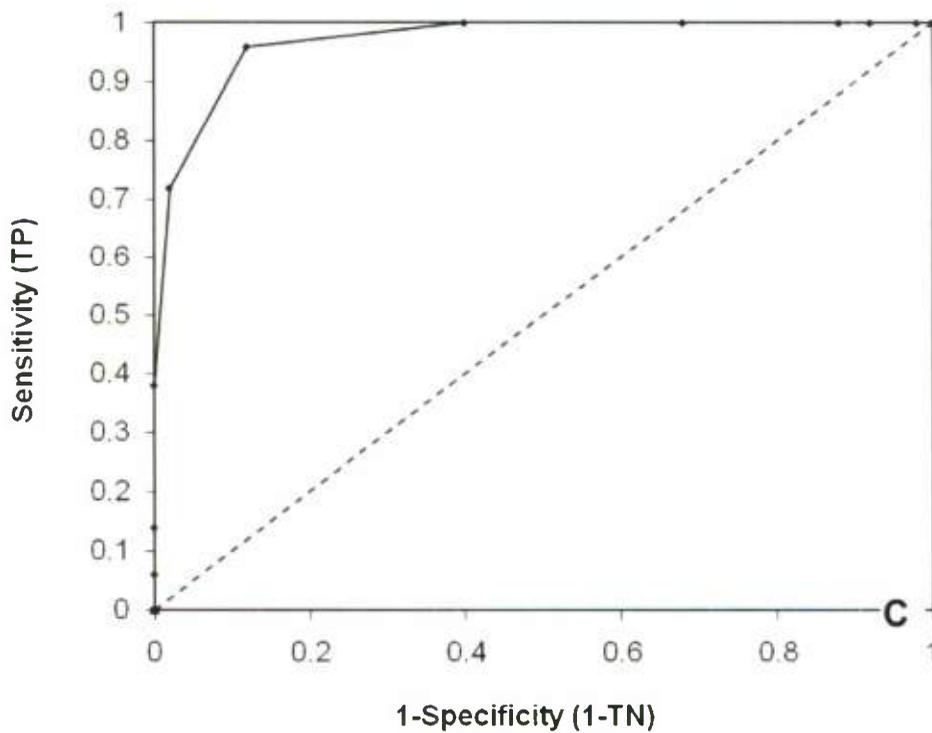
Figure 13C. ROC curve of the data shown in Figure 13B.

ACD vs Rotated Angle (V3,4)
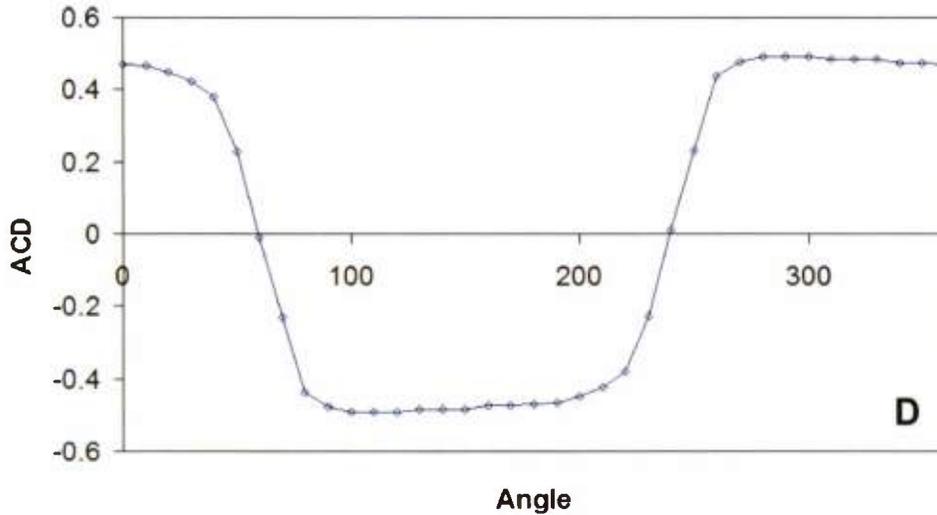
Figure 13D. Plot of angle of rotation of points shown in Figure 13A versus ACD.

Versicolor vs. Virginica @ 280 degree angle
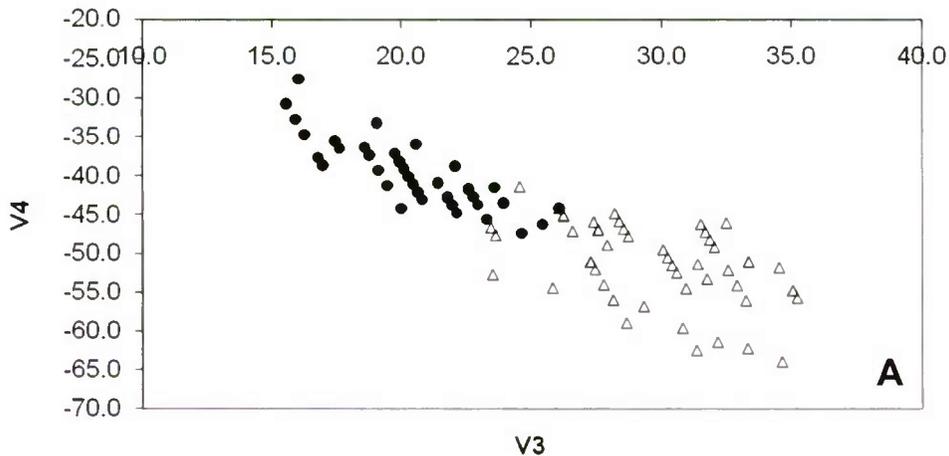
Figure 14A. Petal length (V3) versus petal width (V4) point plot distribution of the 50 replicates for both *I. versicolor* (filled circles) and *I. virginica* (triangles) at 280° rotation of the points.

## Versicolor vs. Virginica @ 280 degree angle
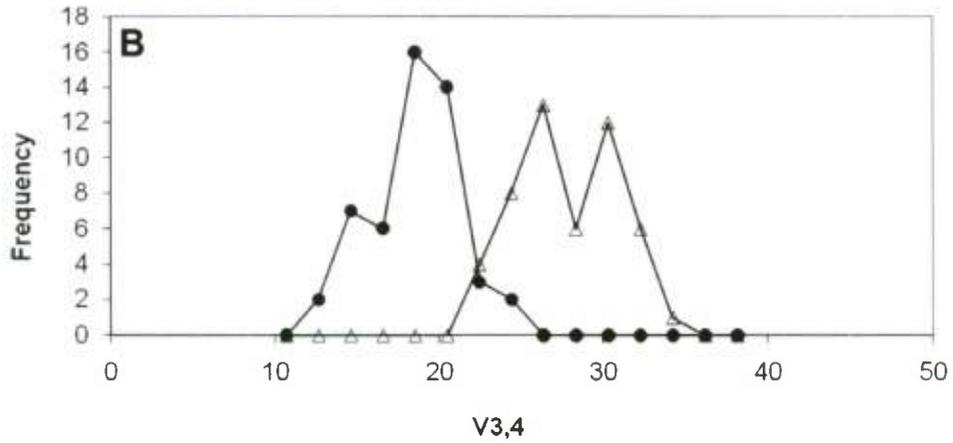
Figure 14B. Petal length (V3) versus petal width (V4) for both *I. versicolor* (filled circles) and *I. virginica* (triangles): frequency distribution at 280° rotation of points.

## Versicolor vs. Virginica @ 280 degree angle
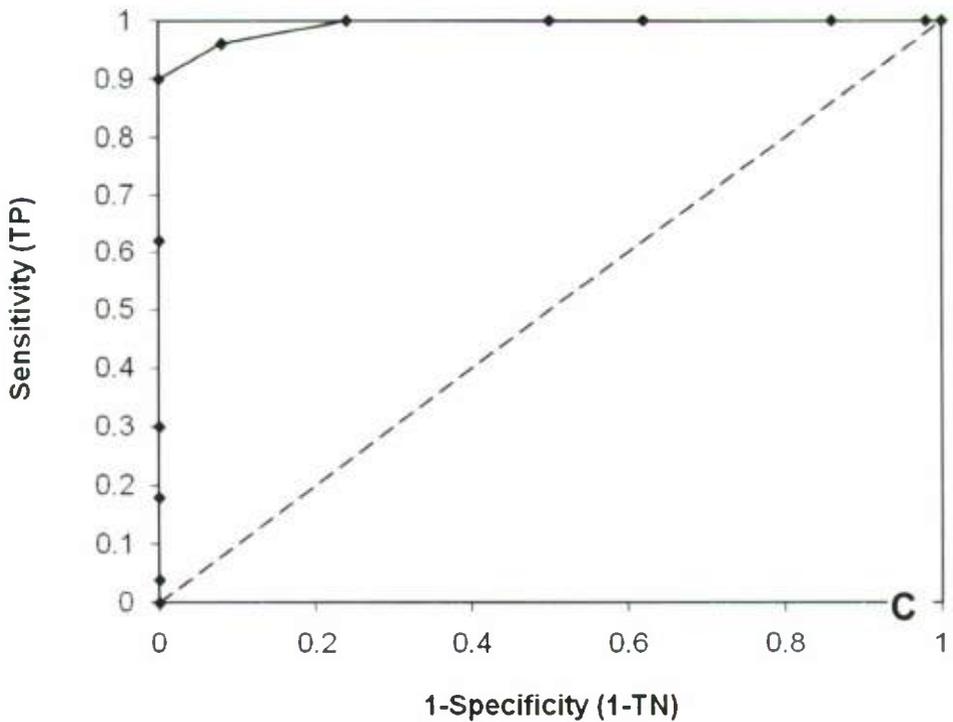
Figure 14C. ROC curve plot of the data shown in Figure 14B.